


Université de Montpellier **Année universitaire 2017-2018**

Rapport de stage de fin d'études

Master 2 de Recherche en Biostatistiques

Université de Montpellier, **Université de Sherbrooke**

Calibration de protocoles d'études de cohorte et calcul de puissance pour l'estimation d'un risque sanitaire en épidémiologie des rayonnements ionisants

Sarra ABAOUBIDA

Soutenu le 04/09/2018

LIEU DU STAGE :	Institut de Radioprotection et de Sûreté Nucléaire (IRSN)
Laboratoire d'acceuil :	Laboratoire d'ÉPIDémiologie des rayonnements ionisants (LEPID)
Maîtres de stage :	Sophie ANCELET (IRSN, LEPID) Éric PARENT (AgroParisTech)
Directeurs de Master :	Benoîte DE-SAPORTA (Université de Montpellier) Éric MARCHAND (Université de Sherbrooke)

Résumé

Au sein du Laboratoire d'ÉPIDémiologie des rayonnements ionisants (LEPID) de l'Institut de Radioprotection et de Sûreté Nucléaire (IRSN), des études de cohorte sont mises en place puis régulièrement mises à jour afin d'estimer les effets sanitaires potentiellement induits par une exposition aux rayonnements ionisants (RI). Ces études de cohorte s'intéressent principalement à des expositions faibles aux RI et à des risques radio-induits associés eux-mêmes potentiellement faibles. Ainsi, à moins d'être suivies relativement longtemps et/ou d'être analysées conjointement avec d'autres cohortes comparables, elles n'ont pas toujours la puissance statistique nécessaire pour mettre en évidence, si elle existe, une association entre la variable réponse d'intérêt (ex., nombre de décès, délai de survenue d'une pathologie donnée) et l'exposition aux RI. Cette limite conduit à un manque de précision fréquent des résultats d'analyse.

Dans ce stage, nous avons proposé une méthode permettant d'approximer la puissance statistique relative à l'utilisation de modèles de survie classiquement utilisés en épidémiologie des RI pour la mise en évidence, si elle existe, d'une association entre un risque sanitaire et une exposition chronique aux RI. Le cas d'étude considéré concerne l'association entre le risque de décès par cancer solide et une exposition chronique et à faibles doses aux rayonnements gamma dans la cohorte des travailleurs d'EDF surveillés pour exposition aux RI. Cette cohorte est actuellement en cours d'extension par le LEPID. Pour un ratio de risque instantané de décès par cancer solide radio-induit de 1.0005 pour une dose de 1 milliSievert - valeur représentative des ordres de grandeurs auxquels s'attendent les épidémiologistes compte-tenu de la littérature internationale - la puissance statistique associée est de 8% si on considère une date de point à 2003 et de 10.2% si on considère une date de point à 2014. Cette puissance est très faible. Cela signifie que, même si un effet radio-induit existe, la probabilité pour que l'étude permette de conclure à l'existence de cet effet est très faible. Aussi, afin d'augmenter la puissance statistique, une première formalisation mathématique du problème d'optimisation d'un protocole d'étude de cohorte a été proposé. Enfin, ce stage a conduit à implémenter un algorithme permettant de simuler des données de survie tronquées à gauche selon un modèle de Cox ou en excès de risque instantané avec covariables dépendantes du temps utilisés en épidémiologie des rayonnements ionisants.

Remerciements

Je tiens à remercier toutes les personnes qui ont contribué au succès de mon stage et qui m'ont aidée lors de la rédaction de ce rapport.

Tout d'abord, j'adresse mes remerciements à mes maîtres de stage, Sophie Ancelet et Éric Parent, pour leurs accueil, le temps passé ensemble et le partage de leurs expertises. Grâce aussi à leur confiance j'ai pu m'accomplir totalement dans mon stage. Sophie a été d'une aide précieuse dans les moments les plus délicats et Éric m'a permis d'acquérir plus de recul sur les statistiques.

Je tiens à remercier vivement Ségolène Bouet-Rivoal et Marion Belloni pour leur aide sur les aspects informatiques.

Je remercie également toute l'équipe du LEPID pour leur accueil chaleureux, leur esprit d'équipe et en particulier Olivier Laurent, qui m'a beaucoup aidée dans la gestion de la base de données.

Enfin, je tiens à remercier toutes les personnes qui m'ont conseillée et relue lors de la rédaction de ce rapport de stage : ma famille, mes amis.

Abréviations

IRSN	**I**nstitut de **R**adioprotection et de **S**ûreté **N**ucléaire
LEPID	**L**aboratoire d'**ÉPID**émiologie
RI	**R**ayonnements **I**onisants
CIRC	**C**entre **I**nternational de **R**echerche sur le **C**ancer
mSv	**m**illi**S**ie**v**ert
EHR	**E**xcess **H**azard **R**atio
TCL	**T**héorème **C**entral **L**imite
BFGS	**B**royden **F**letcher **G**oldfarb

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction

Dans notre quotidien, nous sommes exposés à de nombreuses sources de rayonnement (couramment appelés rayons), visibles ou invisibles. On peut distinguer les rayonnements non-ionisants (e.g., fréquences radio) et les rayonnements ionisants (RI) [Figure 1.1].

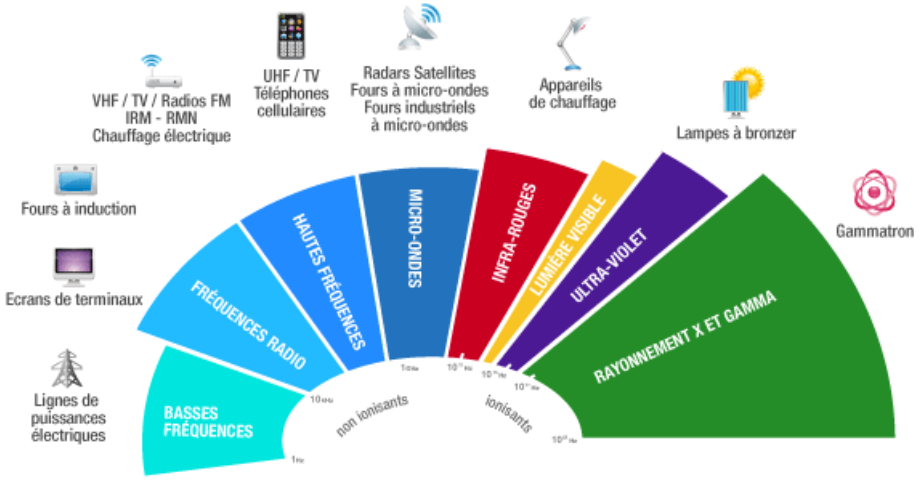

FIGURE 1.1: Les rayonnements dans notre quotidien [2]

Un rayonnement est une émission d'énergie, sous forme d'ondes ou de particules.

Certains rayonnements (ex : X et gamma) sont dit ionisants car ils possèdent des énergies suffisantes pour transformer les atomes qu'ils traversent en ions (un atome qui a perdu ou gagné un ou plusieurs électrons). Le phénomène de radioactivité est ainsi dû à l'instabilité de certains atomes.

Un atome instable va chercher à se stabiliser en émettant différents rayonnements :

— en perdant des protons ou des neutrons ;
— en perdant des particules lourdes, comme par exemple des noyaux d'hélium, dans ce dernier cas on parle de rayonnement alpha ;
— en émettant des photons : rayonnements X et gamma.

Ces phénomènes correspondent au phénomène de radioactivité [32].

L'énergie dégagée n'est pas identique pour tous les rayonnements, et les moyens de s'en protéger sont donc différents. Par exemple, une feuille de papier est suffisante pour arrêter les rayonnements alpha, mais il faut un mètre de béton ou de plomb pour atténuer des rayonnements gamma [Figure 1.2].

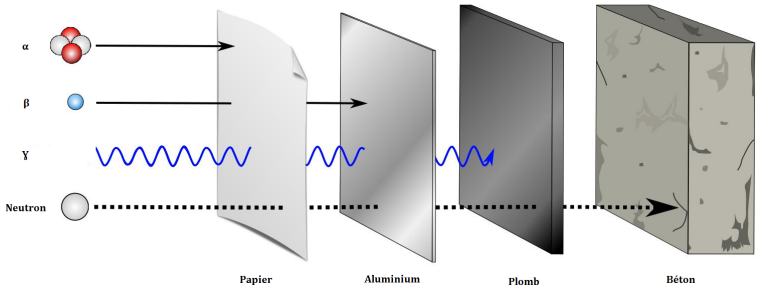

FIGURE 1.2: Le pouvoir de pénétration des différents rayonnements [15]

Nous sommes constamment exposés la radioactivité. En effet, celle-ci peut-être d'origine naturelle car issue de la Terre, du cosmos et de notre alimentation ou d'origine artificielle car associée à la réalisation d'examens ou de thérapies médicales, à des rejets autorisés des installations nucléaires, à des accidents nucléaires ou à des essais d'armes nucléaires dans l'atmosphère.

Deux unités sont fréquemment utilisées pour quantifier une dose reçue de R.I. : le gray (Gy) et le sievert (Sv).

— Le gray (Gy) mesure la dose physiquement « absorbée » par la matière.
— Le sievert (Sv) est l'unité de mesure de la dose équivalente ou efficace, qui permet d'évaluer l'impact du rayonnement sur la matière vivante en tenant compte du type de rayonnement et de la radio-sensibilité des organes ou des tissus [14].

Les RI provoquent des effets biologiques différents sur l'organisme en fonction de la dose reçue. En effet, la dose absorbée par la matière vivante peut entraîner des modifications plus ou moins importantes de celle-ci. Il existe deux types d'effets biologiques :

— les effets déterministes à court terme : une forte irradiation par des RI provoque des effets déterministes à court terme sur les organismes vivants comme, par exemple, des brûlures plus ou moins importantes. Selon la dose et l'organe touché, le délai d'apparition des symptômes varie de quelques heures (nausées, radiodermites) à plusieurs mois. Quand les tissus ne sont pas trop atteints, ces effets sont réversibles et les zones touchées peuvent guérir. Mais, dans le cas d'une très forte irradiation, un trop grand nombre de cellules sont détruites, entraînant la destruction des tissus ou organes irradiés, ce qui peut nécessiter l'amputation d'un membre ou, en cas d'atteinte des systèmes vitaux, peut conduire au décès de la victime ;

— les effets stochastiques à long terme : les expositions à des doses plus ou moins élevées de RI peuvent avoir des effets à long terme sous la forme de cancers et de leucémies. La probabilité d'apparition de ces effets augmente avec la dose : on parle ainsi d'effets stochastiques. Le délai d'apparition après l'exposition est de plusieurs années. Une pathologie radio-induite n'a pas de signature particulière identifiée à ce jour. En effet, on ne connaît pas de marqueur biologique permettant de différencier, par exemple, un cancer pulmonaire dû au tabac, d'un cancer pulmonaire radio-induit.

L'étude des effets sanitaires pouvant être induits par une exposition aux RI est un problème d'intérêt en épidémiologie, en radioprotection et en santé publique. Afin de pouvoir juger de l'existence d'une association entre une exposition aux RI et un risque de maladie, de nombreuses études épidémiologiques (ex., études de cohorte, enquêtes cas-témoin) ont été menées depuis plusieurs décennies.

Ces études ont montré qu'il existe une relation claire entre l'exposition aux RI à des doses fortes et modérées (ex : >100 mSv) et des excès de cancers [32]. Cette relation n'a pas encore été mise en évidence de manière aussi robuste pour de plus faibles doses. Bien que les signaux en ce sens se renforcent pour des expositions de plusieurs dizaines de mSv [23, 21], les effets sur la santé humaine d'une exposition à des doses inférieures à 100 mSv font toujours l'objet de débats scientifiques.

Dans ce stage, on s'intéressera aux travailleurs d'Électricité de France (EDF) qui produit de l'électricité d'origine nucléaire. Dans le cadre de leur activité professionnelle, ces travailleurs sont généralement exposés à des doses de rayonnements gamma cumulées inférieures à 100 mSv.

Une cohorte de travailleurs d'EDF a été mise en place aux cours des années 1990, dans le cadre d'un protocole d'étude internationale coordonnée par le Centre International de Recherche sur le Cancer [6]. Des études épidémiologiques ont été réalisées à partir des données de cette cohorte, afin d'estimer le niveau d'association entre une exposition chronique et à faibles doses aux rayonnements gamma et le risque de décès par cancer. Cependant, les études uniquement basées sur la cohorte EDF n'ont pas pu prouver le caractère statistiquement significatif de cette association ([17], [18], [24]), alors que des études internationales récentes, de plus large ampleur (incluant notamment la cohorte EDF, mais également celles d'autres cohortes françaises, anglaises et américaines de travailleurs du nucléaire) ont pu mettre en évidence une association statistiquement significative entre exposition aux rayonnements gamma et risque de cancer [23].

Ceci engendre les questions suivantes :

- Est-ce-qu'il existe une réelle association entre l'exposition chronique et à faibles doses aux rayonnements gamma et le risque de décès par cancer dans la cohorte de travailleurs d'EDF ?
- Si cet effet existe, pourquoi n'a-t-on pas pu le mettre en évidence dans le cadre de précédentes analyses portant sur cette cohorte ? Est-ce en raison d'un manque de puissance statistique ?

Dans ce travail, nous nous focaliserons uniquement sur le risque de décès par cancer solide.

Un calcul de la puissance statistique actuelle de cette cohorte pour la mise en évidence, s'il existe, d'un excès de risque de décès par cancer solide ainsi que la calibration d'un protocole d'étude de cohorte optimal pour la mise en évidence d'un niveau d'association donné permettent de répondre à la seconde question. En effet, la puissance statistique de la cohorte EDF est sans doute trop faible pour mettre en évidence l'existence d'un effet radio-induit présumé faible.

Sous cette hypothèse, pour augmenter la puissance de l'étude, il faut améliorer le protocole d'étude actuel en déterminant les moyens dont il faudrait disposer (taille de la cohorte, durée de suivie...) pour mettre en évidence un niveau d'association donné. Si, malgré cela, les données recueillies ne permettent pas de retrouver ce niveau d'association alors il serait raisonnable de conclure à la non-existence de l'effet d'intérêt.

Chapitre 2

Contexte

2.1 Cadre du stage

2.1.1 Présentation générale de l'IRSN

L'IRSN est un Établissement Public à caractère Industriel et Commercial (EPIC) créé en 2001. Organisme de recherche et d'expertise, il est le spécialiste public national en matière de recherche et d'expertise sur les risques nucléaires et radiologiques. Placé sous la tutelle conjointe des ministères chargés de la défense, de l'environnement, de l'énergie, de la recherche et de la santé, il concourt aux politiques publiques en matière de sûreté nucléaire et de protection de la santé et de l'environnement au regard des RI. Ses trois missions principales sont :

- La recherche et les services d'intérêt public, incluant l'information du public dans le domaine des risques radiologiques et nucléaires (publication de rapports, organisation d'expositions) ;
- L'appui technique aux autorités publiques compétentes en matière de sûreté nucléaire et de radioprotection pour les activités civiles et de défense ;
- Les prestations d'expertises, d'études et de mesures, pour les organismes publics et privés, nationaux et internationaux.

Les activités de recherche et d'expertise de l'IRSN s'articulent autour de :

- la sûreté nucléaire (réacteurs, cycle du combustible, déchets, applications médicales, etc) ;
- la sûreté des transports de matières radioactives et fissiles ;

— la protection des travailleurs, de la population et de l'environnement contre les risques liés aux rayonnements ionisants ;

— la protection et le contrôle des matières nucléaires ;

— la protection des installations nucléaires et transports de matières radioactives et fissiles contre actes de malveillance.

L'IRSN compte un peu plus de 1800 collaborateurs (ingénieurs, chercheurs, médecins, agronomes, vétérinaires, techniciens) répartis sur 8 sites en France, dont celui de Fontenay-aux-Roses au sein duquel s'est déroulé ce stage de master 2 en biostatistique.

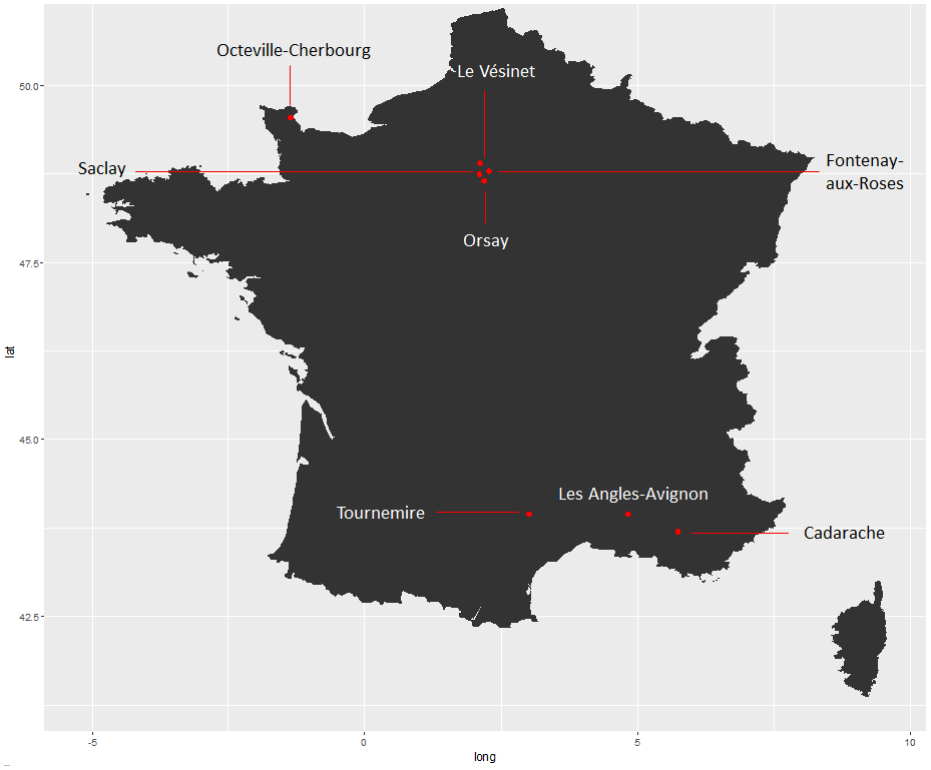

FIGURE 2.1: Situation géographique des différents sites de l'IRSN en France métropolitaine

Les principaux travaux du site de l'IRSN situé à Fontenay-aux-Roses portent sur les domaines suivants :

— expertise nucléaire de défense ;

— radioprotection de l'homme et de l'environnement ;

— crises et interventions ;

— sûreté des installations nucléaires ;

— sûreté de la gestion des déchets ;

— sûreté des transports.

2.1.2 Présentation du LEPID

Le stage s'est déroulé au sein du Laboratoire d'ÉPIDémiologie des rayonnements ionisants (LEPID) de l'IRSN. Créé en 1991, le LEPID fait partie du Service de recherche sur les Effets biologiques et SANitaires des rayonnEments ionisants (SESANE) de la direction Santé du pôle Santé et Environnement.

L'objectif principal des travaux de recherche du LEPID est d'améliorer les connaissances sur les effets sanitaires associés à une exposition aux RI chez l'homme, notamment dans le cadre d'expositions chroniques et à faibles doses. Dans cette optique, le LEPID met en place et assure le suivi épidémiologique de cohortes d'individus dont les expositions aux RI sont d'origine environnementale, médicale ou professionnelle. Ces travaux visent à évaluer la validité des hypothèses sous-jacentes au système de radioprotection actuel, élaboré au niveau international pour gérer la protection de différentes populations exposées (patients, travailleurs, population générale), et contribuent à son évolution.

Au delà de cet objectif principal, le LEPID cible trois axes de recherche spécifiques, présentant un intérêt majeur en radioprotection :

- les effets sanitaires des contaminations internes, notamment chroniques, aux RI. Ceux-ci sont en effet beaucoup moins connus que les effets des expositions externes. En situation de contamination interne et contrairement aux expositions externes, l'irradiation continue après la fin de l'exposition du fait de l'incorporation par ingestion ou inhalation de radionucléides et les organes sont exposés de manière hétérogène ;
- les effets sanitaires des expositions médicales (radiographies, scanners) et environnementales (radioactivité naturelle) à faibles doses pendant l'enfance. En effet, peu d'études ont été menées à ce jour dans ce contexte. Seules quelques études indiquent que la radiosensibilité serait plus élevée à dose fixée si la dose a été reçue pendant l'enfance plutôt qu'à l'âge adulte ;
- les risques de pathologies non-cancéreuses radio-induites (maladies cardio-vasculaires, cataractes...). En effet, bien que plusieurs études récentes suggèrent un accroissement de la fréquence de certaines pathologies non cancéreuses à des niveaux de dose relativement faibles, davantage d'études sont nécessaires afin de valider ou pas ces premiers résultats sur d'autres populations et ainsi d'améliorer la connaissance

des effets des RI sur les maladies cardio-vasculaires.

Par ailleurs, le LEPID dispose d'un axe de recherche transverse, relatif, d'une part, à la gestion des bases de données et, d'autres part, à l'application et au développement de méthodes statistiques adaptées pour la modélisation de données épidémiologiques complexes, l'estimation de risques sanitaires radio-induits ou encore pour la prise en compte de sources d'incertitude potentielles dans le cadre de l'estimation ou de la prédiction de risques sanitaires radio-induits.

Enfin, le LEPID s'est toujours questionné sur les problèmes de puissance statistique et de calibration optimale de protocoles d'étude de cohorte auxquels il est confronté régulièrement, d'où l'intérêt de ce stage.

2.2 Position du problème

En épidémiologie des RI, une approche standard utilisée pour augmenter la puissance statistique des études de cohorte est de construire des cohortes internationales combinant les données épidémiologiques issues de plusieurs cohortes nationales afin d'augmenter le nombre d'individus. Une autre approche consiste à mettre à jour régulièrement les cohortes existantes afin d'augmenter le temps de suivi des individus. Enfin, les calculs de puissance statistique sont réalisés dans un cadre fréquentiste avec, pour objectif spécifique, d'évaluer la capacité des données épidémiologiques disponibles à mettre en évidence, s'il existe, un accroissement de risque d'une ampleur supposée chez les sujets exposés d'une cohorte par rapport à des sujets non exposés [5]. Néanmoins, à notre connaissance, aucun outil statistique n'est actuellement disponible pour :

a) calculer la puissance statistique d'une étude de cohorte à nombre de sujets et temps de suivi préfixés, dans le contexte spécifique des modèles dose-réponse standards dans le domaine ;

b) calibrer un protocole d'étude de cohorte (ex., nombre de sujets, temps de suivi...) permettant d'atteindre une puissance minimale fixée (typiquement 80%) ou optimale sous contrainte, quand l'objectif spécifique est de mettre en évidence, si elle existe, une association entre une variable réponse et une exposition aux RI en univers incertain (e.g., statuts vitaux, taux de base, valeurs des facteurs de risque inconnus) et à partir de

modèles dose-réponse standards dans le domaine (e.g., régression de Poisson, modèle de survie en excès de risque).

Bien que les travaux de Little et al. (2010) [19] aient récemment abordé le problème, ceux-ci se focalisent uniquement sur la question a), ne considèrent pas les modèles dose-réponse classiquement utilisés en épidémiologie des RI (e.g., régression de poisson, modèles de survie...) et se placent dans un cadre statistique fréquentiste ne permettant pas de tenir compte de l'incertitude sur les nombreuses quantités intervenant dans la calibration des protocoles d'étude.

Enfin, bien que la calibration de plans d'expérience optimaux pour la mise en place d'études expérimentales (ex., essais cliniques) ait fait l'objet de nombreux travaux de recherche ces dernières années[3, 9], cela est nettement moins le cas pour la calibration de protocoles d'étude pour données d'exposition longitudinales telles que celles rencontrées en épidémiologie des RI, venant complexifier les problèmes. Quelques logiciels [33] ont été proposés pour la calibration de plans d'expérience optimaux dans ce contexte mais ces derniers concernent essentiellement la recherche clinique pour laquelle les variables du plan d'expérience sont contrôlables expérimentalement. Or en épidémiologie des RI, les mesures d'exposition ainsi que les valeurs des différents facteurs de risque susceptibles d'entrer en jeu dans l'occurrence de la pathologie d'intérêt ne sont pas contrôlables par l'épidémiologiste. Par ailleurs, contrairement aux études expérimentales, un certain nombre de facteurs incontrôlables et inhérents à tout processus observationnel peuvent venir affaiblir le protocole d'étude initial (ex., perdus de vue, données manquantes, sources de biais ...). Ainsi, des travaux de recherche semblent nécessaires afin de développer des outils statistiques mieux appropriés à ces contextes.

2.3 Objectifs du stage

Les objectifs du stage sont de développer puis de mettre en oeuvre sur les données de la cohorte des travailleurs EDF et en se focalisant sur l'association entre une exposition chronique et à faibles doses aux rayonnements gamma et le risque de décès par cancer solide :

- une méthode permettant de calculer la puissance statistique relative à une étude de cohorte professionnelle quand l'objectif spécifique est de mettre en évidence, s'ils existent, des risques radio-induits ;

- une méthode permettant de calibrer les composantes-clés du protocole d'étude de cohorte (ex., combien de sujets ? Quel temps de suivi ?...) afin d'atteindre une puissance minimale fixée (typiquement 80%) ou optimale sous contrainte.

2.4 Intérêts du stage

Ce travail de stage présente deux intérêts principaux :

- un intérêt direct pour les épidémiologistes du LEPID. En effet, il doit au moins permettre d'apporter de premiers éléments de réponse aux questions a) et b) auxquelles ces derniers sont régulièrement confrontés en amont de la mise en place d'études de cohorte ou de la mise à jour de cohortes existantes.

- un intérêt en santé publique et en radioprotection en contribuant indirectement à l'amélioration des connaissances sur les effets sanitaires radio-induits à faibles doses. Il s'agit de fournir des informations concernant la faisabilité de la mise en place d'études de cohortes ayant une puissance statistique optimale, compte-tenu de contraintes budgétaires fixées (par exemple), pour la mise en évidence de ces risques potentiellement faibles comme le souligne le Comité Scientifique des Nations Unies sur les effets des rayonnements ionisants (UNSCEAR) : « there is a need to recognize that studies of high statistical power are necessary in order to be sure that health effects at these doses have not been missed » (UNSCEAR, Rapport 2012).

Chapitre 3

Cas d'étude : la cohorte EDF

3.1 Historique de la cohorte EDF

Électricité de France produit de l'électricité d'origine nucléaire depuis 1961. Dans le cadre d'un projet concu par le CIRC visant à analyser la mortalité des travailleurs de l'industrie nucléaire de différents pays, EDF a constitué au cours des années 1990 une cohorte incluant des agents surveillés pour exposition aux RI.[17]. Ces travaux ont ensuite été poursuivis par le laboratoire d'épidémiologie de l'IRSN et le sont encore aujourd'hui.

3.2 Description générale de la cohorte (suivi et dosimétrie)

La cohorte EDF se compose de 30425 agents statutaires. On désignera par **cohorte initiale** la cohorte contenant les travailleurs EDF suivis de 1961 à 2003. Dans le but d'augmenter le nombre d'années de suivi des travailleurs inclus dans la cohorte initiale, une extension sur la période de 2003 à 2014 de la base de données épidémiologiques pré-existante est en cours. La cohorte mise à jour sera appelée **cohorte étendue**.

Les conditions d'inclusion des agents statutaires dans l'étude sont :

- travailler durant au moins un an au sein de l'entreprise ;
- faire l'objet d'une surveillance pour exposition aux RI pour la première fois entre 1961 (année de démarrage de la production d'électricité d'origine nucléaire au sein d'EDF) et le 31 décembre 2003 [24].

Ces critères d'inclusion sont identiques à ceux utilisés dans les études précédentes de la cohorte EDF [24].

Tous les agents statutaires ayant travaillé au moins un an dans l'entreprise et ayant été surveillés pour exposition aux RI entre 1961 et 2003 ont été inclus dans l'étude.

3.2.1 Suivi

La date d'entrée dans l'étude a été définie pour chaque agent comme la date la plus récente parmi la date de première surveillance dosimétrique, la date de premier emploi plus un an et le 1er janvier 1968 [24].
La date de fin de suivi (en d'autres termes, la date de sortie de l'étude) a été définie pour chaque agent comme la date la plus ancienne parmi le 31 décembre 2003 pour la cohorte initiale (respectivement le 31 décembre 2014 pour la cohorte étendue), la date de décès et la date de dernières nouvelles pour les perdus de vue.

3.2.2 Dosimétrie

Les agents statutaires inclus dans la cohorte EDF sont principalement exposés par voie externe à des rayonnements gamma. Une partie de ces agents est également exposée par voie externe aux rayonnements neutroniques. Des contaminations internes peuvent survenir en cas de situations accidentelles, toutefois celles-ci sont rares [17].

L'exposition des agents aux rayonnements gamma est notamment mesurée à partir des résultats de la surveillance réglementaire individuelle réalisée à partir de dosimètres portés sur la poitrine et lus mensuellement [17].

On dispose des doses de rayonnements gamma des travailleurs inclus dans la cohorte EDF estimées annuellement entre 1961 et 2003 pour la cohorte initiale et entre 1961 et 2014 pour la cohorte étendue.

Comme les expositions gamma sont majoritaires et que les incertitudes relatives à la dosimétrie neutron sont importantes [30], on s'est focalisé sur les rayonnements gamma dans ce stage.

3.3 Analyse descriptive

Parmi les 30425 travailleurs de la cohorte, 93,6% sont des hommes. La durée moyenne de suivi des agents est de 15,8 ans [17]. Ces caractéristiques sont présentées de manière détaillée pour la cohorte initiale dans le tableau 3.1.

TABLE 3.1: Caractéristiques générales de la cohorte initiale.

	Nombre (%)
Démographie	
Effectif d'agents statutaires	30425
Hommes	28467 (93,56 %)
Femmes	1958 (6,44 %)
Durée moyenne de suivie en années	15,8
Age moyen en fin de suivi en années	45,5
Dosimétrie gamma	
Effectifs d'agents présentant des doses gamma >0	25671 (84,37 %)
Dose (en mSV) cumulée gamma moyenne	17,07
Dosimétrie neutron	
Effectifs d'agents présentant des doses neutron >0	4159 (13,67 %)
Dose (en mSV) cumulée neutron moyenne	0,21
Statut vital au 31 Décembre 2003	
Vivants	29467 (96,85 %)
Perdus de vue	66 (0,22 %)
Décédés	892 (2,93 %)
par cancers	311 (1,02 %)
par cancers solides	295 (0,97 %)
par leucémies	16 (0,05 %)

Le tableau ci-dessous représente les âges moyen, minimal et maximal à l'entrée dans l'étude, à la sortie de l'étude (date de point) et au décès par cancer solide (on s'interesse uniquement au cancer solide dans ce stage), caractérisant la cohorte EDF.

TABLE 3.2: Ages caractérisant la cohorte EDF.

	Moyenne	**Minimum**	**Maximum**
Ages à l'entrée dans l'étude	29.48	17.56	69.03
Ages à la sortie de l'étude	45.78	19.09	97.02
Ages au décès par cancer solide	54.55	30.7	89.4

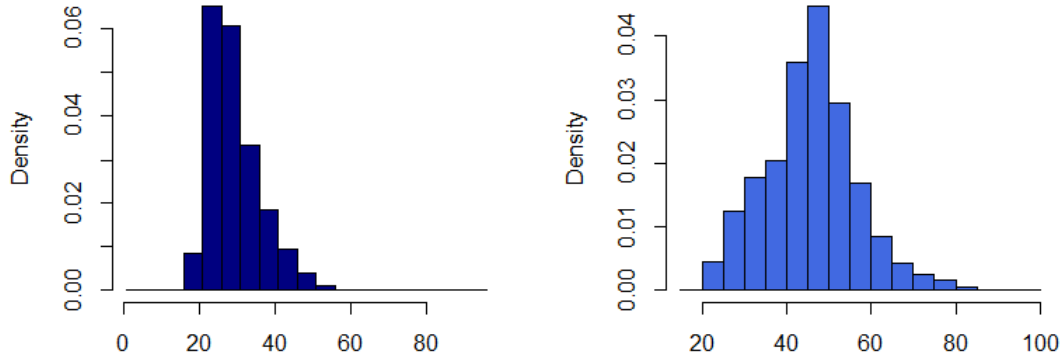

FIGURE 3.1: Histogramme des âges à l'entrée dans l'étude (à gauche, en bleu foncé) et à la sortie de l'étude (à droite, en bleu clair) (date de point 2003).

D'après l'histogramme des âges des travailleurs inclus dans la cohorte EDF à leur sortie de l'étude (date de point 2003), on remarque que la population d'étude est jeune ce qui explique pourquoi seulement 3% des travailleurs de la cohorte étaient décédés en 2003. En particulier, on observe seulement 295 décès par cancer solide à la date de point en 2003 ce qui représente seulement 1% de la population de la cohorte. Le phénomène du travailleur sain est observé dans la plupart des cohortes jeunes. Cela se traduit par le fait que la mortalité chez les travailleurs de la cohorte EDF soit inférieure à la mortalité de la population française [17, 25, 5].

La figure ci-dessous représente la distribution des âges de décès par cancer solide dans la cohorte EDF initiale.

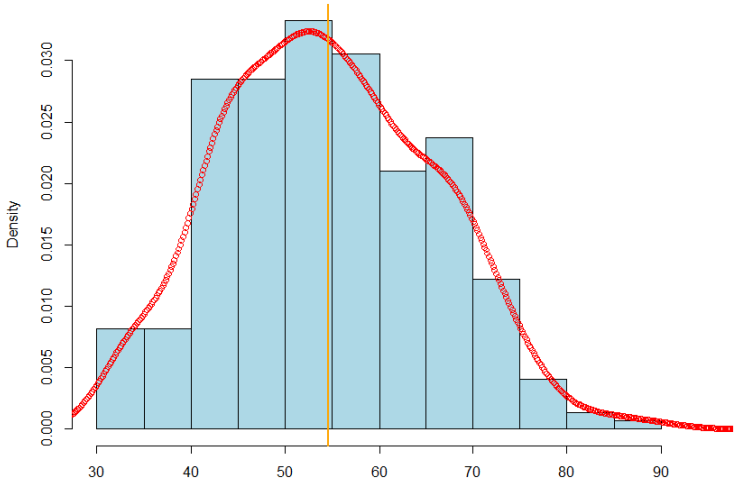

FIGURE 3.2: Histogramme des âges de décès par cancer solide dans la cohorte EDF initiale.

Chapitre 4

Modélisation et simulation de données de survie avec covariables temps-dépendantes

Dans ce chapitre, nous décrivons tout d'abord deux classes de modèles de survie possibles pour décrire les données de survie observées dans les cohortes professionnelles étudiées en épidémiologie des RI. Les modèles décrits dans ce chapitre seront utilisés pour les calculs de puissance statistique et l'optimisation de protocoles d'étude, présentés dans les chapitres suivants. Les modèles sont présentés dans le cas spécifique des données de la cohorte EDF mais cette description peut aisément être adaptée à d'autres cohortes. Quelques généralités sur les modèles de survie sont rappelées dans l'annexe A.1.

Les données de survie disponibles dans les cohortes professionnelles étudiées en épidémiologie des RI présentent trois particularités : a) elles sont tronquées à gauche à l'âge d'entrée dans l'étude des individus ; b) elles sont censurées à droite et c) elles dépendent de covariables d'exposition aux RI qui varient avec le temps (comme expliqué ci-dessous). Dans ce chapitre, nous décrivons une méthode permettant de simuler de telles données de survie.

4.1 Description des modèles

Soit X_i l'âge (en jours) au décès par cancer solide du travailleur $i \in \{1, \ldots, n\}$. Considérant

l'âge des individus comme échelle de temps, ces temps de décès sont ainsi tronqués à gauche à l'âge d'entrée dans la cohorte de chaque individu, noté r_i par la suite. Soit C_i l'âge (en jours) de censure à droite du travailleur i (également tronqué à gauche en r_i) qui peut correspondre à trois événements différents :

- décès dû à une autre cause qu'un cancer solide (censure aléatoire) ;
- perte de vue du travailleur (ex. changement d'entreprise) (censure aléatoire) ;
- survie jusqu'à la date de fin de suivi (31 décembre 2003 pour la cohorte initiale ou 31 décembre 2014 pour la cohorte étendue) (censure déterministe).

Les variables aléatoires X_i et C_i sont supposées indépendantes.

Pour chaque travailleur i, on observe donc :

- un temps de survie $T_i = min(X_i, C_i)$, tronqué à gauche en r_i
- un indicateur binaire de non censure $\delta_i = \mathbb{1}_{X_i \leq C_i}$

Dans ce travail, les temps de censure aléatoire C_i ont été supposés uniformément distribués dans le temps.

On souhaite modéliser l'association entre le risque de décès par cancer solide et l'exposition cumulée aux rayonnements gamma, à partir de temps de survie T_i tronqués à gauche et d'indicateurs binaires de non-censure δ_i. Dans le cas d'un cancer solide, on suppose un temps de latence de 10 ans. En d'autres termes, on adopte l'hypothèse communément admise selon laquelle un cancer solide ne peut être considéré comme radio-induit qu'après un délai minimal de 10 ans après l'exposition.

Pour chaque travailleur i, le temps de décès par cancer solide X_i est supposé suivre un modèle de survie dont le risque instantané associé $h_i(t)$ au temps t est défini par :

$$h_i(t) = h_0(t)\rho(D_i^{cum}(t-10), \beta)$$

où $h_0(t)$ est le risque instantané de base et la fonction ρ désigne le ratio de risques instantanés $\frac{h_i(t)}{h_0(t)}$. Cette dernière décrit la forme de l'association entre la dose cumulée aux rayonnements gamma du travailleur i à l'âge t laggée de 10 ans (en milliSievert (mSv)), notée $D_i^{cum}(t-10)$, et le risque instantané total de décès par cancer solide $h_i(t)$.

Dans la suite, on considérera deux fonctions ρ classiquement utilisées en épidémiologie des R.I :

— **Celle associée au modèle de Cox :**

$\rho(D_i^{cum}(t-10), \beta) = \exp(\beta D_i^{cum}(t-10))$

— **Celle associée au modèle en Excès de Risque Instantané (EHR) :**

$\rho(D_i^{cum}(t-10), \beta) = 1 + \beta D_i^{cum}(t-10)$

Dans ces deux modèles, β représente le degré d'association entre le risque de décès par cancer solide et la dose cumulée gamma. On peut remarquer que le modèle en EHR correspond en fait à un développement limité au premier ordre au voisinage de 0 du modèle de Cox (en d'autres termes, si β est proche de 0, alors les deux modèles sont très similaires).

Habituellement, dans un modèle de Cox, on ne précise pas la forme du risque de base $h_0(t)$, car seule la vraisemblance partielle est utilisée pour l'estimation fréquentiste du coefficient de risque inconnu d'intérêt β. C'est d'ailleurs pour cette raison que le modèle de Cox est souvent qualifié de modèle semi-paramétrique. Néanmoins, dans le contexte spécifique d'un calcul de puissance statistique ou de calibration d'un protocole d'étude de cohorte, nous verrons par la suite qu'il est indispensable d'être capable de disposer d'un modèle complet permettant de générer des temps de décès. Il est donc nécessaire de choisir une forme (paramétrique) pour le risque instantané de base $h_0(t)$. Dans ce stage, nous avons choisi une fonction $h_0(t)$ constante par morceaux, définie par :
$\forall t \in I_l =]c_{l-1}, c_l]$ et pour $l \in \{1, \ldots, L\}$

$$h_0(t) = \lambda_l$$

où $0 = c_0 < c_1 < \ldots < c_L$ est une partition donnée du temps.

La loi de probabilité des âges au décès X_i peut être définie par l'une des cinq fonctions équivalentes suivantes :

Sa fonction de survie :

La fonction de survie de X_i est, pour tout $t \geq 0$, la probabilité de survivre jusqu'à l'instant t, c'est-à-dire

$$S_i(t) = P(X_i \geq t)$$

Sa fonction de répartition :

La fonction de répartition de X_i représente, pour tout t, la probabilité de décéder avant

l'instant t, c'est-à-dire

$$F_i(t) = P(X_i < t) = 1 - S_i(t)$$

Sa densité de probabilité :

C'est la fonction $f_i(t) \geq 0$ telle que pour tout $t \geq 0$

$$F_i(t) = \int_0^t f_i(u)du$$

Si la fonction de répartition F_i admet une dérivée au point t alors

$$f_i(t) = \lim_{h \to 0} \frac{\mathbb{P}(t \leq X_i < t + h)}{h} = F_i'(t) = -S_i'(t)$$

Pour t fixé, la densité de probabilité s'interprète comme la probabilité de décéder dans un intervalle de temps infinitésimal après l'instant t.

Sa fonction de risque instantané h_i (ou taux de hasard) :

Le risque instantané au temps t quantifie la probabilité de décéder dans un intervalle de temps infinitésimal après t, conditionnellement au fait d'avoir survécu jusqu'au temps t :

$$h_i(t) = \lim_{h \to 0} \frac{\mathbb{P}(t \leq X_i < t + h | X_i \geq t)}{h} = \frac{f_i(t)}{S_i(t)}$$

On remarque que

$$h_i(t) = \frac{-S_i'(t)}{S_i(t)} = -ln(S_i(t))'$$

4.2 Simulation de données de survie tronquées à gauche avec covariables dépendantes du temps

Comme nous le verrons dans le chapitre suivant, être capable de simuler des données de survie tronquées à gauche et avec covariables d'exposition dépendantes du temps - telles que celles de la cohorte des travailleurs EDF - est indispensable au calcul de la puissance statistique dans un objectif de mise en évidence de potentiels risques sanitaires radio-induits.

Or, la simulation de données de survie tronquées à gauche et avec covariables dépendantes du temps, selon les modèles de Cox ou en EHR précédemment décrits, n'est pas triviale. A notre connaissance, peu de littérature existe à ce sujet. Ainsi, ce travail de stage a tout d'abord consisté à adapter la méthode de simulation proposée par Hendry et al. (2014) [11] au contexte des données de cohortes professionnelles rencontrées en épidémiologie des R.I (e.g.,cohorte EDF). Le but de cette partie est de décrire la méthode de simulation proposée. Pour cela, il faut simuler des temps de décès X_i et de censure C_i pour chaque travailleur i $(1, \ldots, n)$ puis en déduire les temps de survie $T_i = min(X_i, C_i)$ et l'indicateur de non censure δ_i. Pour rappel, dans ce travail, les temps de censure C_i ont été supposés uniformément distribués dans le temps.

4.2.1 Simulation des temps de décès X_i : principe général

Nous commençons par décrire le principe général de la méthode de simulations dont la preuve de fonctionnement repose sur le théorème ci-après.

Soient $\mathcal{S} = \{s_0, s_1, s_2, ..., s_J\}$ une partition du temps avec $0 = s_0 < s_1 < ... < s_J$ et $Z(t)$ une covariable dépendante du temps (ex., exposition aux rayonnements gamma d'un travailleur EDF) mais supposée constante sur chaque intervalle $I_j =]s_{j-1}, s_j] : Z(t) = Z_j$ pour tout $s_{j-1} < t \leq s_j$. Notons $\gamma_j = \exp(Z_j\beta)$, le ratio de risques instantanés sur I_j. Soit g une fonction telle que :

$$g(0) = 0,\ g(t) \text{ croissante pour } t > 0 \text{ et } g^{-1}(t) \text{ est différentiable}$$

Théorème : [11]
Soit E une variable aléatoire qui suit une loi exponentielle par morceaux d'intensités γ_j sur $]g^{-1}(s_{j-1}), g^{-1}(s_j)]$ pour $j \in \{1, \ldots, J\}$. La densité de probabilité de E est alors donnée par :

$$k_E(t) = \prod_{h=1}^{j-1} \exp\big(-\gamma_h(g^{-1}(s_h) - g^{-1}(s_{h-1}))\big) \times \gamma_j \exp\big(-\gamma_j(t - g^{-1}(s_{j-1}))\big) \times \mathbb{1}_{\{g^{-1}(s_{j-1}) < t \leq g^{-1}(s_j)\}}$$

Soit Y une version tronquée de E sur le support $[g^{-1}(a)), g^{-1}(b)]$, de densité :

$$f_Y(t) = \frac{k_E(t)\mathbb{1}_{\{g^{-1}(a)\leq t\leq g^{-1}(b)\}}}{K_E(g^{-1}(b)) - K_E(g^{-1}(a))}$$

où $K_E(t)$ désigne la fonction de répartition correspondante à la loi de E.
Alors $g(Y)$ suit un modèle de Cox tronqué sur le support $[a, b]$, avec covariable dépendante du temps et un taux de base défini par :

$$h_0(t) = \frac{d}{dt}[g^{-1}(t)]$$

Preuve : (cf. annexe A.2.1)

Remarques :

- Ce théorème est aussi valable pour le modèle en EHR, en prenant $\gamma_j = 1 + \beta Z_j$. La démonstration reste la même.
- La méthode d'acceptation-rejet (cf. annexe A.2.2) permet facilement de générer une variable aléatoire qui suit une loi exponentielle par morceaux tronquée, à partir d'une variable qui suit une loi exponentielle par morceaux. En effet, pour $y \in]g^{-1}(a), g^{-1}(b)[$, la constante $M = \frac{1}{K_E(g^{-1}(b))-K_E(g^{-1}(a))} > 0$ est alors telle que :

$$f_Y(y) \leq M \times k_E(y)$$

4.2.2 Algorithme de simulation de temps de survie $T_i = min(X_i, C_i)$

L'algorithme générique de simulation de temps de survie $T_i = min(X_i, C_i)$, selon un modèle de Cox tronqué sur le support [a,b], avec risque instantané de base défini par $h_0(t) = \frac{d}{dt}[g^{-1}(t)]$ et covariable dépendante du temps est décrit ci-dessous :

1. Définir une fonction g telle que $g(0) = 0$, $g(t)$ croissante pour $t > 0$ et $g^{-1}(t)$ différentiable
2. Définir une partition du temps $\mathcal{S} = \{s_0, s_1, s_2, ..., s_J\}$ avec $s_0 = 0$
3. Définir les bornes de troncature a et b des temps de survie à simuler
4. Choisir une valeur de paramètre β
5. Pour $i \in \{1, \ldots, n\}$
   (a) Calculer $\{\gamma_{ij}\}_{j=1}^{J} = \{\exp(Z_{ij}\beta)\}_{j=1}^{J}$ avec Z_{ij} la valeur de covariable spécifique à l'individu i sur l'intervalle I_j

(b) Générer Y_i selon une loi exponentielle par morceaux tronquée avec intensités $\gamma_{i1}, \ldots, \gamma_{iJ}$ aux points de changement de temps $g^{-1}(s_1), \ldots, g^{-1}(s_J)$ et bornes de troncature $g^{-1}(a)$ et $g^{-1}(b)$.

(c) Calculer le temps de décès $X_i = g(Y_i)$

6. Générer le temps de censure C_i selon une loi uniforme entre a et b

7. Calculer le temps de survie $T_i = min(X_i, C_i)$ et l'indicateur de non censure δ_i.

Le même algorithme peut être utilisé pour la simulation de données de survie selon un modèle en EHR tronqué sur le support [a,b], avec risque instantané de base défini par $h_0(t) = \frac{d}{dt}[g^{-1}(t)]$ pour tout t et covariable dépendante du temps. Il suffit de calculer $\{\gamma_{ij}\}_{j=1}^{J} = \{1 + \beta Z_{ij}\}_{j=1}^{J}$ à l'étape 5 (a).

L'application du théorème d'Hendry a été adaptée au cas du choix d'une fonction de risque instantané de base constante par morceaux (cf. chapitre 5, section 5.3.1)

4.3 Inférence fréquentiste d'un modèle de survie avec covariables dépendantes du temps

Comme nous le verrons dans le chapitre suivant, le calcul de la puissance statistique d'intérêt pour la mise en évidence de risques sanitaires radio-induits, à partir des modèles de Cox ou en EHR avec troncature à gauche et covariables dépendantes du temps précédemment décrits, nécessite de mener l'inférence fréquentiste, par maximum de vraisemblance, de ces modèles.

Il n'existe pas de formule analytique pour l'estimateur du maximum de vraisemblance $\hat{\beta}$ et pour l'écart-type d'estimation associé $\hat{\sigma}$ dans le cadre des modèles de Cox et en EHR.

Par ailleurs, les fonctions *coxph* et *phreg* disponibles sous R ne semblent pas fonctionner correctement pour l'estimation par maximum de vraisemblance d'un modèle de Cox avec covariables dépendantes du temps, à partir de données de survie tronquées à gauche. Aussi, afin d'estimer le coefficient de risque β, un algorithme d'estimation par maximum de vraisemblance, basé principalement sur la méthode de Broyden Fletcher Goldfarb Shanno (BFGS), a été implémenté sous R via la fonction *optim*.

L'algorithme BFGS est un algorithme de minimisation (et donc également de maximisation) d'une fonction basé sur des directions de descente. L'idée principale de cette méthode

est d'éviter de devoir calculer explicitement la matrice hessienne de la fonction à minimiser en utilisant plutôt une approximation de celle-ci, construite à partir du calcul de différents gradients successifs. L'utilisation de l'algorithme BFGS a nécessité d'implémenter en R la vraisemblance des modèles considérés ainsi que les gradients associés. Leurs expressions sont détaillées en annexe (A.3,A.4).

Chapitre 5

Puissance statistique pour la mise en évidence d'un effet sanitaire radio-induit à partir d'un modèle de survie

Soit un modèle de survie (e.g., Cox ou EHR) permettant de relier l'âge de survie des travailleurs jusqu'au décès par cancer solide à leurs expositions respectives aux rayonnements gamma. Un test statistique paramétrique peut être utilisé pour mettre en évidence, s'il existe, un effet sanitaire radio-induit. Suite au calcul d'une statistique de test à partir de données collectées, une réponse est apportée, grâce aux méthodes de statistique inférentielle, à la question d'intérêt suivante :

— Compte-tenu des données disponibles, peut-on rejeter l'hypothèse H_0 : "Absence d'association entre une exposition aux RI et une variable réponse sanitaire" ?

Un risque fixé α dit, de première espèce, quantifie la probabilité de conclure à tort à l'existence d'association (i.e., rejeter H_0). Un risque β dit, de deuxième espèce, quantifie la probabilité de ne pas rejeter l'hypothèse H_0 alors qu'une telle association existe. $1 - \beta$ est appelé "la puissance du test".

Cette démarche de test statistique peut conduire à passer sous silence un certain nombre de questions :

— Lorsque l'association radio-induite est concrètement importante, on imagine bien qu'il faut moins d'observations pour la mettre en évidence que lorsqu'elle est petite ... mais combien au juste ? A-t-on les moyens, en termes de nombre de mesures, de mettre en évidence l'association recherchée ? Faut-il s'y prendre autrement et changer le protocole d'étude de cohorte ? Ces questions seront abordées dans le chapitre suivant.

5.1 Notions générales et exemple Gaussien

5.1.1 Généralités

Soient $X_1, \ldots, X_n$ n variables aléatoires indépendantes et identiquement distribuées (i.i.d) dont la loi de probabilité dépend d'un vecteur de paramètres inconnus $\theta \in \mathscr{V} \subset \mathbb{R}^p$. Soient $\mathscr{V}_0$ et $\mathscr{V}_1$ deux ensembles de $\mathscr{V}$, tels que $\mathscr{V}_0 \cap \mathscr{V}_1 = \varnothing$. Un test statistique est une méthode permettant de trancher entre deux hypothèses sur le vecteur de paramètres inconnus θ, au vu des données observées :

$$H_0 : \theta \in \mathscr{V}_0 \qquad \text{contre} \qquad H_1 : \theta \in \mathscr{V}_1$$

L'hypothèse nulle H_0 est celle qu'on cherche à vérifier. Si le résultat d'échantillonnage conduit au rejet de l'hypothèse nulle alors la conclusion acceptée dans ce cas s'appelle l'hypothèse alternative, notée H_1.

Une statistique $U = h(X_1, \ldots, X_n)$, fonction des variables aléatoires observables, est appelée statistique de test si elle joue le rôle de variable de décision. Idéalement, cette statistique de test doit apporter le maximum d'informations sur le problème posé et sa loi de probabilité doit être connue au moins sous H_0. Par ailleurs, elle n'a aucune raison d'être unique et le choix entre plusieurs statistiques candidates est une question difficile : un test fondé sur une première statistique peut conclure au non rejet de l'hypothèse nulle tandis que la considération d'une autre statistique peut conclure à son rejet. Enfin, la statistique de test choisie et sa distribution sous H_0 dépend généralement de la loi de probabilité des variables aléatoires observables $X_1, \ldots, X_n$ et de la taille de l'échantillon observé.

On appelle région de rejet, région critique ou zone d'acceptation W toute région de $\mathbb{R}^n$ contenant l'ensemble des réalisations de $X_1, \ldots, X_n$ conduisant au rejet de l'hypothèse nulle :

$$W = \{(x_1, \ldots, x_n) \in \mathbb{R}^n / \text{on rejette } H_0\}$$

où $x_1, \ldots, x_n$ est une réalisation de $X_1, \ldots, X_n$.

Lorqu'on dispose d'une statistique de test U et de sa loi sous H_0, on peut écrire :

$$W = \{u \in \mathbb{R} / \text{on rejette } H_0\}$$

où u est une réalisation de U pour un échantillon observé $x_1, \ldots, x_n$. La détermination de la région de rejet W dépend de l'erreur de première espèce α. Elle s'obtient en résolvant l'équation : $\mathbb{P}_{H_0}(\text{rejeter } H_0) = \mathbb{P}_{H_0}(W) = \alpha$.

La puissance d'un test statistique est définie comme la probabilité de rejeter l'hypothèse nulle à raison c'est-à-dire sous l'hypothèse alternative H_1. Formellement, elle est définie comme suit : puissance $= \mathbb{P}_{H_1}(\text{rejeter } H_0) = \mathbb{P}_{H_1}(W)$

TABLE 5.1: Risques d'erreur lors de la prise de décision

Vérité	H_0 est vraie	H_1 est vraie
Décision		
On choisit H_0	1-α	β
On choisit H_1	α	1-β =puissance

On considère généralement que la puissance doit être au moins égale à 0.8 pour être satisfaisante.

5.1.2 L'exemple Gaussien : calcul exact de la puissance statistique

Comme cela est le cas dans l'exemple décrit dans cette section, il est parfois possible de calculer analytiquement la puissance statistique associée à un test paramétrique.

Soient $X_1, \ldots, X_n$, n variables aléatoires i.i.d qui suivent une $\mathcal{N}(\mu, \sigma)$ avec σ le paramètre d'écart-type supposé connu.
On veut tester :

$$H_0 : \mu = \mu_0 \qquad \text{contre} \qquad H_1 : \mu > \mu_0$$

La statistique de test choisie est la moyenne empirique de l'échantillon : $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. La loi de $\bar{X}$ est connue sous H_0 : $\bar{X} \sim \mathcal{N}(\mu_0, \frac{\sigma}{\sqrt{n}})$.

La zone de rejet W est définie comme l'ensemble des réalisations de $X_1, \ldots, X_n$ telles que la moyenne empirique associée s'éloigne significativement de H_0 : $W = \{\bar{X} > \text{seuil}\}$. Le seuil de définition de la zone de rejet W est calculable analytiquement :

$$\begin{aligned}
\mathbb{P}_{H_0}(W) = \alpha &\Leftrightarrow \mathbb{P}_{H_0}(\bar{X} > \text{seuil}) = \alpha \\
&\Leftrightarrow \mathbb{P}_{H_0}(\bar{X} \leq \text{seuil}) = 1 - \alpha \\
&\Leftrightarrow \mathbb{P}_{H_0}(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \leq \frac{\text{seuil} - \mu_0}{\sigma/\sqrt{n}}) = 1 - \alpha \\
&\Leftrightarrow \mathbb{P}_{H_0}(\mathcal{N}(0,1) \leq \frac{\text{seuil} - \mu_0}{\sigma/\sqrt{n}}) = 1 - \alpha
\end{aligned}$$

donc

$$\frac{\text{seuil} - \mu_0}{\sigma/\sqrt{n}} = z_\alpha \Leftrightarrow \text{seuil} = \mu_0 + z_\alpha \times \frac{\sigma}{\sqrt{n}}$$

où z_α est le quantile à $(1-\alpha)\%$ d'une $\mathcal{N}(0,1)$. Ainsi, la région de rejet est $W = \{\bar{X} > \mu_0 + z_\alpha \times \frac{\sigma}{\sqrt{n}}\}$

De même, dans cet exemple, la puissance du test est calculable analytiquement. Néanmoins, pour cela, il est indispensable de choisir une alternative ponctuelle pour H_1, c'est à dire de se fixer une valeur pour le paramètre d'expérance μ sous H_1, car il est nécessaire de définir la loi de la statistique de test $\bar{X}$ sous H_1.On se place donc dans le cadre du test suivant :

$$H_0 : \mu = \mu_0 \qquad \text{contre} \qquad H_1 : \mu = \mu_1$$

où $\mu_1 > \mu_0$.

Sous l'hypothèse alternative, $\bar{X} \sim \mathcal{N}(\mu_1, \frac{\sigma}{\sqrt{n}})$. La puissance du test considéré vaut alors :

$$\begin{aligned}
\mathbb{P}_{H_1}(\bar{X} > \mu_0 + z_\alpha \times \frac{\sigma}{\sqrt{n}}) &= 1 - \mathbb{P}_{H_1}(\bar{X} \leq \mu_0 + z_\alpha \times \frac{\sigma}{\sqrt{n}}) \\
&= 1 - \mathbb{P}_{H_1}\left(\frac{\bar{X} - \mu_1}{\frac{\sigma}{\sqrt{n}}} \leq \frac{\mu_0 + z_\alpha \times \frac{\sigma}{\sqrt{n}} - \mu_1}{\frac{\sigma}{\sqrt{n}}}\right) \\
&= 1 - \Phi\left(\frac{\mu_0 + z_\alpha \times \frac{\sigma}{\sqrt{n}} - \mu_1}{\frac{\sigma}{\sqrt{n}}}\right)
\end{aligned}$$

où Φ est la fonction de répartition d'une loi normale $\mathcal{N}(0,1)$.

Cet exemple permet de mettre en évidence que la puissance d'un test statistique dépend de l'erreur de première espèce α, de la taille n de l'échantillon observé, des paramètres μ et σ et de la région critique W.

La figure 5.1 réprésente la courbe de puissance associée au test statistique considéré en fonction de la valeur de μ_1 et pour des valeurs fixées de μ_0, σ, n et α.

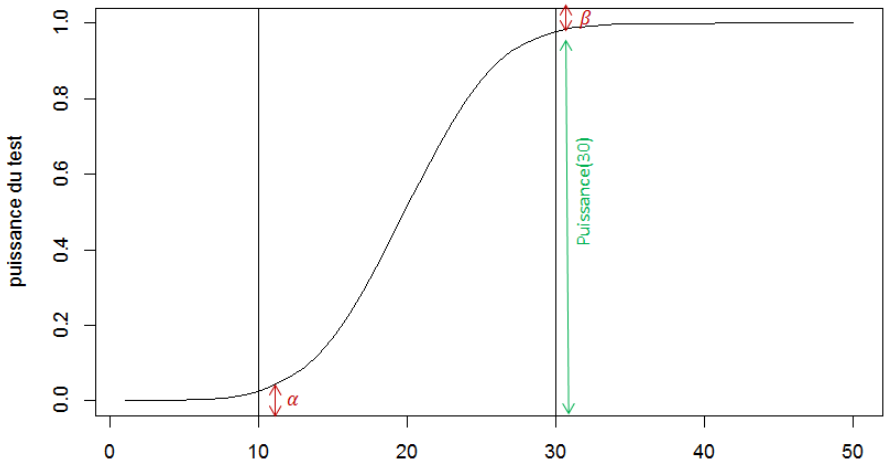

FIGURE 5.1: Puissance du test en fonction de μ_1, pour $\mu_0 = 10$, $\sigma = 50$, $n = 100$ et $\alpha = 0.05$

5.2 Approximation de la puissance statistique pour les modèles de survie

Dans le cadre de l'analyse des données de la cohorte EDF (cf. chapitre 2), on cherche à mettre en évidence, si elle existe, une association entre le risque de décès par cancer solide et une exposition chronique et à faibles doses aux rayonnements gamma. Les modèles de Cox et en EHR décrits dans le chapitre 3 font partie des modèles probabilistes utilisés en épidémiologie des R.I pour décrire cette association à partir de données de survie tronquées à gauche et avec covariables d'exposition dépendantes du temps telles que celles de la cohorte des travailleurs EDF.

Pour rappel, pour ces deux modèles, le ratio entre le risque instantané total de décès par cancer solide et le risque instantané de base est défini par :

— Modèle de Cox :

$$\exp(\beta \times D_i^{cum}(t-10))$$

— Modèle en EHR :

$$1 + \beta \times D_i^{cum}(t - 10)$$

On peut donc interpréter le paramètre $\exp(\beta)$ du modèle de Cox (respectivement $1 + \beta$ pour le modèle en EHR) comme le ratio de risque instantané par cancer solide pour une dose de 1 mSv. Une absence d'effet sanitaire radio-induit sur le risque de décès par cancer solide se traduit donc formellement par l'égalité suivante : $\beta = 0$. Au contraire, l'existence d'un effet sanitaire radio-induit se traduit par : $\beta \neq 0$. Le test paramétrique à considérer prend donc la forme suivante :

$$H_0 : \beta = 0 \qquad \text{contre} \qquad H_1 : \beta \neq 0$$

La statistique de Wald définie dans ce cas par le ratio : $U = \frac{\hat{\beta}}{\hat{\sigma}}$ avec $\hat{\beta}$ l'estimateur du maximum de vraisemblance de β et $\hat{\sigma}$ l'écart-type d'estimation de β associé, est classiquement utilisée pour ce type de test. Néanmoins, contrairement à la statistique de test de l'exemple de la section 5.1.2, la loi de la statistique de test U sous H_0 est uniquement connue asymptotiquement c'est à dire quand la taille de l'échantillon n tend vers $+\infty$. Dans ce cas et sous des conditions de régularité, la statistique U est consistante c'est-à-dire qu'elle tend en loi vers une $\mathcal{N}(0, 1)$ sous H_0.

Par analogie avec l'exemple Gaussien de la section 5.1.2, la région de rejet W du test ci-dessus s'écrit sous la forme $W = \{|U| > seuil\}$. Le seuil du test peut donc uniquement être approché asymptotiquement pour une valeur d'erreur de première espèce α fixée :

$$\begin{aligned}
\mathbb{P}_{H_0}(\text{rejeter } H_0) = \alpha &\Leftrightarrow \mathbb{P}_{H_0}(|U| > \text{seuil}) = \alpha \\
&\Leftrightarrow 2 \times \mathbb{P}_{H_0}(U > \text{seuil}) \simeq \alpha \qquad \text{(hypothèse normalité asymptotique)} \\
&\Leftrightarrow \mathbb{P}_{H_0}(U \leq \text{seuil}) \simeq 1 - \frac{\alpha}{2}
\end{aligned}$$

En posant $z_{\frac{\alpha}{2}}$ le quantile à $(1 - \frac{\alpha}{2})\%$ d'une $\mathcal{N}(0, 1)$, on obtient : seuil $\simeq z_{\frac{\alpha}{2}}$.

Comme dans l'exemple précédent, le calcul de la puissance statistique associée à ce test nécessite de se fixer une valeur ponctuelle de β sous l'hypothèse alternative H_1. Par ailleurs, comme le seuil du test peut uniquement être approché asymptotiquement, il en va de même de la puissance statistique du test qui peut être approchée par :

$$\begin{aligned}
\mathbb{P}_{H_1}(\text{rejeter } H_0) &\simeq \mathbb{P}_{H_1}(|U| > z_{\frac{\alpha}{2}}) \\
&= \mathbb{P}_{H_1}\left(\left|\frac{\hat{\beta}}{\hat{\sigma}}\right| > z_{\frac{\alpha}{2}}\right) \\
&= \mathbb{P}_{H_1}\left(|\hat{\beta}| > z_{\frac{\alpha}{2}} \times \hat{\sigma}\right) \\
&= \mathbb{P}_{H_1}\left(\hat{\beta} \in]-\infty, -z_{\frac{\alpha}{2}} \times \hat{\sigma}[\cup]z_{\frac{\alpha}{2}} \times \hat{\sigma}, +\infty[\right) \\
&= \mathbb{E}_{H_1}\left(\mathbb{1}_{\{\hat{\beta} \in]-\infty, -z_{\frac{\alpha}{2}} \times \hat{\sigma}[\cup]z_{\frac{\alpha}{2}} \times \hat{\sigma}, +\infty[\}}\right)
\end{aligned}$$

A ce stade, une deuxième difficulté apparaît car, dans les modèles de Cox et en EHR, la loi de $\hat{\beta}$ sous H_1 n'a pas une forme analytique connue. On peut néanmoins contourner ce problème en calculant une approximation de Monte-Carlo de la puissance statistique d'intérêt.

Considérons N réalisations indépendantes de la cohorte EDF et $\hat{\beta}_i$ (i=1,...,N) l'ensemble des N estimateurs du maximum de vraisemblance associés du paramètre β et d'écart-type d'estimation respectif $\hat{\sigma}_i$. Alors, d'après la loi forte des grands nombres :

$$\begin{aligned}
\mathbb{E}_{H_1}\left(\mathbb{1}_{\{\hat{\beta} \in]-\infty, -z_{\frac{\alpha}{2}} \times \hat{\sigma}[\bigcup]z_{\frac{\alpha}{2}} \times \hat{\sigma}, +\infty[\}}\right) &\simeq \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{\{\hat{\beta}_i \in]-\infty, -z_{\frac{\alpha}{2}} \times \hat{\sigma}_i[\bigcup]z_{\frac{\alpha}{2}} \times \hat{\sigma}_i, +\infty[\}} \\
&= \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{\{0 \in]-\infty, \hat{\beta}_i - z_{\frac{\alpha}{2}} \times \hat{\sigma}_i[\bigcup]\hat{\beta}_i + z_{\frac{\alpha}{2}} \times \hat{\sigma}_i, +\infty[\}} \\
&= \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{\{0 \notin]\hat{\beta}_i - z_{\frac{\alpha}{2}} \times \hat{\sigma}_i, \hat{\beta}_i + z_{\frac{\alpha}{2}} \times \hat{\sigma}_i[\}}
\end{aligned}$$

Finalement, il vient :

$$\text{Puissance} \simeq \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{\{0 \notin]\hat{\beta}_i - z_{\frac{\alpha}{2}} \times \hat{\sigma}_i, \hat{\beta}_i + z_{\frac{\alpha}{2}} \times \hat{\sigma}_i[\}}$$

Un estimateur sans biais de la variance de cet estimateur de Monte-Carlo de la puissance statistique est donné par :

$$\frac{1}{N}\Big[\frac{1}{N-1}\sum_{i=1}^{N}\big(k(\hat{\beta}_i) - \frac{1}{N}\sum_{j=1}^{N} k(\hat{\beta}_j)\big)^2\Big]$$

avec $k(\hat{\beta}_i) = \mathbb{1}_{\{0 \notin]\hat{\beta}_i - z_{\frac{\alpha}{2}} \times \hat{\sigma}_i, \hat{\beta}_i + z_{\frac{\alpha}{2}} \times \hat{\sigma}_i[\}}$.

Si les $\hat{\beta}_i$ sont i.i.d, les $k(\hat{\beta}_i)$ le sont aussi et donc, grâce au théorème limite central, on peut définir un intervalle de confiance pour l'estimateur de la Puissance (i.e., $\overline{k(\hat{\beta}_i)}$) pour un certain niveau de test fixé α.

Dans le cas d'un modèle de Cox ou en EHR, la puissance statistique associée au test

$$H_0 : \beta = 0 \qquad \text{contre} \qquad H_1 : \beta \neq 0$$

peut ainsi être approchée par simulations, en suivant la démarche générale ci-dessous :

1. Simuler les temps de survie, tronqués à gauche et avec covariables dépendantes de temps, relatifs à N cohortes de 30425 travailleurs EDF selon un modèle de Cox ou en EHR et en se plaçant sous H_1, c'est à dire pour une valeur ponctuelle $\beta_1 \neq 0$.
2. Pour chaque jeu de données simulées i=1,...,N :
   — Ajuster un modèle de Cox ou en EHR par maximum de vraisemblance afin d'obtenir $\hat{\beta}_i$ et $\hat{\sigma}_i$
   — Pour une erreur de première espèce α fixée (e.g., $\alpha = 0.05$), calculer l'intervalle de confiance à $(1-\alpha)\%$ (e.g., 95%) suivant : $]\hat{\beta}_i - z_{\frac{\alpha}{2}} \times \hat{\sigma}_i, \hat{\beta}_i + z_{\frac{\alpha}{2}} \times \hat{\sigma}_i[$
3. Calculer la proportion d'intervalles de confiance à $(1-\alpha)\%$ ne contenant pas 0 : ce taux de couverture approxime la puissance du test au niveau α.

5.3 Application à la cohorte des travailleurs EDF

5.3.1 Simulation de temps de survie par cancer solide

Pour approximer la puissance statistique de la cohorte des travailleurs EDF pour la mise en évidence d'un risque de décès par cancer solide radio-induit, une première étape a été de simuler des données de survie par cancer solide, similaires à celles de la cohorte EDF. Pour cela, nous avons utilisé les données d'exposition aux rayonnements gamma disponibles pour tous les travailleurs de la cohorte initiale (date de point 2003) et de la cohorte étendue (date de point 2014), ainsi que leur âge respectif à l'entrée dans la cohorte.

Dans le cadre de ce stage, l'algorithme de simulation de temps de survie décrit dans la section 4.2.2 a été adapté a) au cas des données de la cohorte des travailleurs EDF et b) au cas de modèles de survie avec taux de base constant par morceaux

Choix de la partition de temps et des bornes de troncature :
On a choisi une partition du temps annuelle commune à tous les travailleurs c'est à dire allant de $s_0 = 0$ à $s_J = 100$ ans et par pas de 1 an (soit 100 intervalles I_j). Les âges aux décès par cancer solide simulés sont tronqués à gauche à l'âge d'entrée dans l'étude du travailleur i, noté r_i. La borne de troncature à gauche des données simulées varie donc selon le travailleur. Néanmoins, on supposera les données de survie non tronquées à droite i.e., $b = +\infty$ pour tout travailleur i. Finalement, le temps de survie simulé pour chaque travailleur i appartient à $[r_i, +\infty[$.

Choix des valeurs du coefficient de risque de décès par cancer solide β :
Les valeurs du coefficient de risque de décès par cancer solide radio-induit β choisies pour les calculs de puissance statistique ont été proposées par les épidémiologistes du LEPID à partir de la littérature scientifique internationale [20, 23] : a) de 1×10^{-4} à 9×10^{-4} par pas de 2×10^{-4}, b) de 1×10^{-3} à 9×10^{-3} par pas de 2×10^{-3} et c) de 1.1×10^{-2} à 2.1×10^{-2} par pas de 2×10^{-3}.

Choix de la fonction g :
Dans le cadre de ce stage, l'algorithme de simulation de temps de survie décrit dans la section 4.2.2 a été légèrement adapté au cas d'une fonction de risque instantané de base h_0 constante par morceaux. (cf. chapitre 4)

Les valeurs de paramètres λ_l et les points de discontinuité c_l de h_0 utilisés pour la simulation de données ont été choisis à partir des données de la cohorte EDF initiale, en estimant les λ_l sur une partition de temps à pas équidistants de 5 ans, puis en regroupant les intervalles pour lesquels les λ_l étaient proches (on obtient la valeur du nouveau λ_l en faisant la moyenne sur les anciens). Cela nous a permis de définir une fonction de risque instantané de base constante sur les 3 intervalles de temps suivants : $\lambda_1 = 2.44 \times 10^{-7}$ sur l'intervalle $]0, 40]$, $\lambda_2 = 2.48 \times 10^{-6}$ sur $]40, 60]$ et $\lambda_3 = 4.87 \times 10^{-5}$ sur $]60, +\infty]$. La figure 5.2 représente le risque instantané de base de décès par cancer solide correspondant. Ce choix de taux de base traduit le fait que plus un travailleur est agé, plus la probabilité de décéder augmente. Par ailleurs, le choix d'une telle partition de temps permet de s'assurer que, pour chaque jeu de données simulé, au moins un

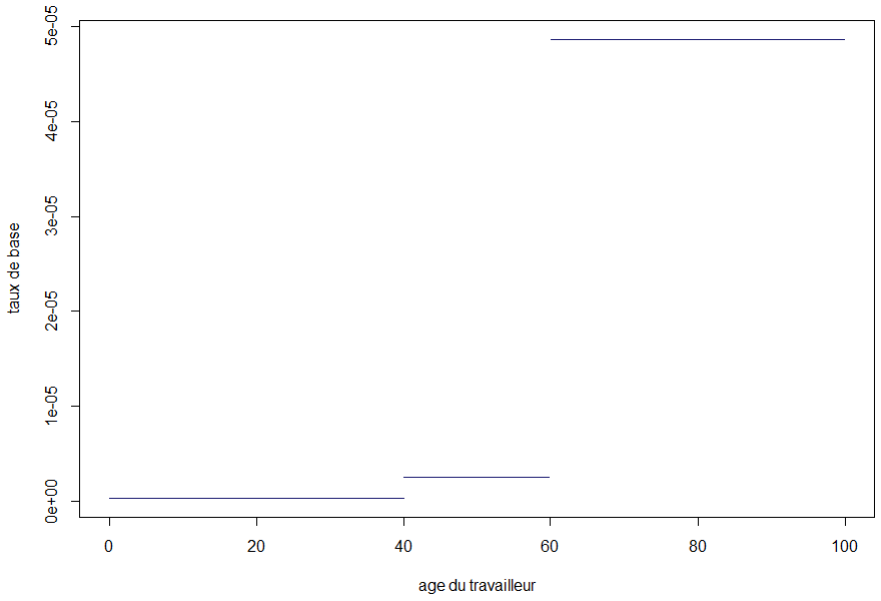

FIGURE 5.2: Risque instantané de base de décès par cancer solide en fonction de l'âge, utilisé pour la simulation de données de survie

travailleur décédera par cancer solide dans chaque intervalle de temps, une condition indispensable à l'estimation de tous les paramètres λ_l.

On souhaite que $h_0(t) = \lambda_l$ pour tout $t \in]c_{l-1}, c_l]$. Or, d'après l'algorithme de simulation générique décrit dans la section 4.2.2, $h_0(t) = \dfrac{dg^{-1}(t)}{dt}$, donc, dans ce cas, $g(t) = \dfrac{t}{\lambda_l}\mathbb{1}_{\{t\in]c_{l-1},c_l]\}}$.

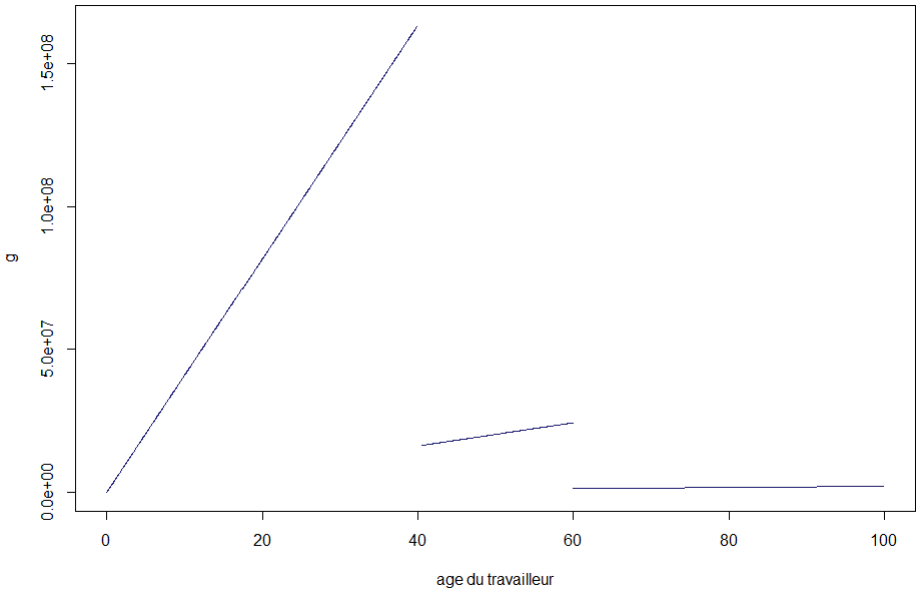

FIGURE 5.3: Fonction $g(t) = \dfrac{t}{\lambda_l}\mathbb{1}_{\{t\in]c_{l-1},c_l]\}}$ en fonction de l'âge

Or, la Figure 5.3 montre que cette fonction g n'est pas croissante (croissante par morceaux) ce qui ne permet pas d'appliquer directement le théorème d'Hendry décrit dans la section 4.2.1. Nous avons ainsi légèrement adapté la méthode de simulation proposée au cas d'un taux de base constant par morceaux en posant $g(t) = t$ pour tout t et en calculant $\gamma_{ij} = \lambda_{l_j} exp(\beta Z_{ij})$ pour le modèle de Cox ou $\gamma_{ij} = \lambda_{l_j}(1 + \beta Z_{ij})$ pour le modèle en EHR avec λ_{l_j} la valeur du taux de base sur l'intervalle de temps I_j. La preuve du théorème décrit dans la section 4.2.1 reste valide pour de telles intensités. Ainsi, on a simulé les âges aux décès par cancer solide selon une loi exponentielle par morceaux de points de changement $\{1, \ldots, 100\}$ et d'intensités $\{\gamma_{ij}\}_{j=1}^{J}$ definies ci-dessus avec $\lambda_{l_j} = \lambda_1 \ \forall l_j \in]0, 40]$, $\lambda_{l_j} = \lambda_2 \ \forall l_j \in]40, 60]$ et $\lambda_{l_j} = \lambda_3 \ \forall l_j \in]60, +\infty[$.

Quelques précisions algorithmiques :
Pour simuler des âges au décès par cancer solide Y_i (non tronqués) selon une loi exponentielle par morceaux, d'intensités $\{\gamma_{ij}\}_{j=1}^{J}$ définies ci-dessus, et de points de changement $\{1, \ldots, 100\}$, on a utilisé la fonction *rpexp* du package *msm* de R.

Pour simuler des âges aux décès par cancer solide X_i tronqués à gauche en r_i, plusieurs âges aux décès non tronqués Y_i ont été simulés pour chaque travailleur i pour en garder un seul, supérieur à r_i (le premier par exemple). Ceci est équivalent à la méthode acceptation rejet.

L'âge à la censure aléatoire du travailleur i a été simulé selon une loi uniforme entre son âge d'entrée dans l'étude r_i et s_J. Cette censure aléatoire représente le fait que le travailleur i peut soit a) décéder d'une autre cause que le cancer solide ou b) être perdu de vue. Il existe également une censure déterministe, fixée à la date de point de l'étude. L'âge à la censure C_i du travailleur i est ainsi défini comme l'âge minimum entre son âge à la censure aléatoire et son âge à la date de point de l'étude (ex., 2003 pour la cohorte initiale).

Finalement, on a calculé le minimum entre X_i et C_i pour obtenir l'âge de survie T_i de chaque travailleur i ainsi que l'indicateur de non-censure associé δ_i.

Graphiques de données simulées pour certaines valeurs de β :
Les figures 5.4, 5.5 et 5.6 représentent les temps de décès X_i, temps de censure C_i et temps de survie T_i simulés selon le modèle de Cox pour une cohorte de 30425 travailleurs et pour trois valeurs possibles du coefficient de risque β de décès par cancer solide

radio-induit : 0.011, 0.007, 0.0007. Ces simulations ont été effectuées pour $\lambda_1 = 2.44 \times 10^{-7}$, $\lambda_2 = 2.48 \times 10^{-6}$, $\lambda_3 = 4.87 \times 10^{-5}$ et pour la date de point en 2003.

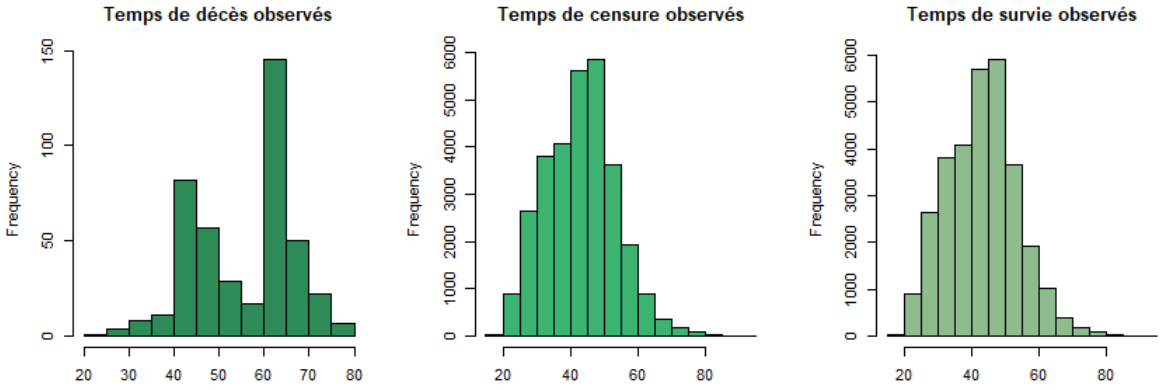

FIGURE 5.4: (De gauche à droite) Temps de décès, temps de censure et temps de survie simulés pour 30425 travailleurs quand $\beta = 0.011$

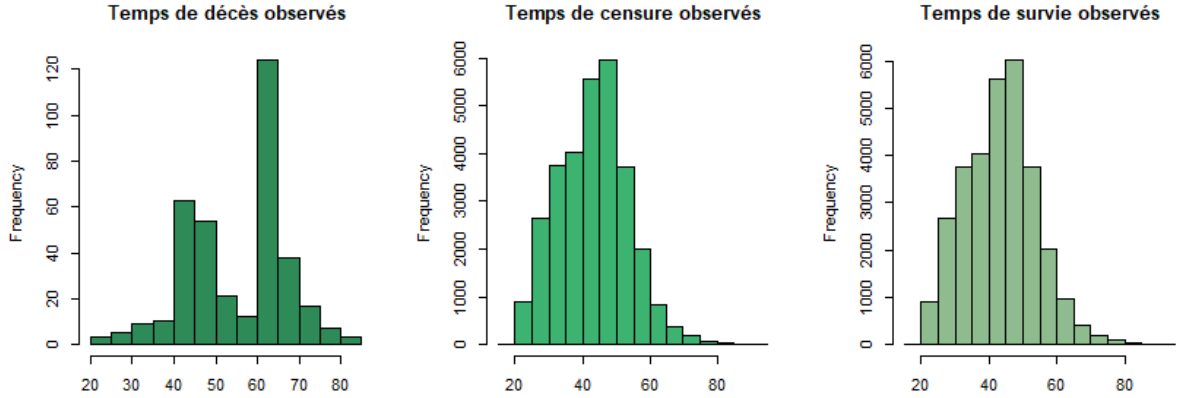

FIGURE 5.5: (De gauche à droite) Temps de décès, temps de censure et temps de survie simulés pour 30425 travailleurs quand $\beta = 0.007$

On remarque que :

— La distribution des temps de survie T_i simulés est fortement influencée par la distribution des temps de censure C_i (aléatoire et déterministe). Ceci est normal car très peu de décès par cancer solide sont observés dans la cohorte (1.42 % de décès par cancer solide pour la simulation selon $\beta = 0.011$, 1.2 % pour $\beta = 0.007$ et 1.08 % pour $\beta = 0.0007$).

— La distribution des temps de décès est fortement influencée par le modèle constant par morceaux choisi pour décrire le risque instantané de base de décès par cancer solide (points de changement à 40 et 60 ans). Elle ressemble finalement assez peu aux âges de décès par cancer solide observés dans la cohorte EDF (cf. Figure 3.2)

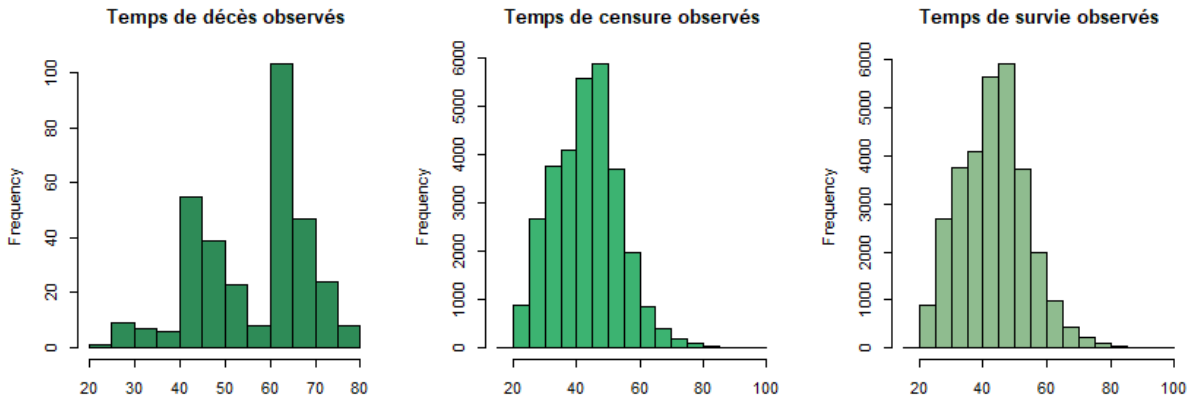

FIGURE 5.6: (De gauche à droite) Temps de décès, temps de censure et temps de survie simulés pour 30425 travailleurs quand $\beta = 0.0007$

5.3.2 Inférence par maximum de vraisemblance

L'estimation par maximum de vraisemblance des paramètres des modèles de Cox ou en EHR considérés pour le cas d'étude EDF - c'est-à-dire $(\beta, \lambda_1, \lambda_2, \lambda_3)$ - a été réalisée via la fonction *optim* du package *stats* de R, qui minimise une fonction selon un algorithme adapté. Les arguments utilisés pour cela sont :

- par : paramètres initiaux (l'algorithme converge plus rapidement si les paramètres initiaux se rapprochent des vrais paramètres) ;
- fn : l'opposé de la fonction de vraisemblance (car on cherche à maximiser la vraisemblance) ;
- gr : l'opposé du gradient de la vraisemblance ;
- method : BFGS.

En ce qui concerne l'estimation des écarts types des estimateurs des paramètres obtenus, la fonction *fdHess* du package *nlme* de R a été utilisée pour l'approximation de la hessienne de la vraisemblance. La matrice obtenue approxime la matrice d'information de Fisher observée. Il suffit donc de l'inverser puis de calculer les racines de sa diagonale pour obtenir les écarts types de $(\hat{\beta}, \hat{\lambda}_1, \hat{\lambda_2}, \hat{\lambda_3})$.

Enfin, on a choisi une erreur de première espèce $\alpha = 0.05$ pour le calcul des intervalles de confiance de β puis les approximations de puissance statistique d'intérêt.

5.3.3 Résultats

Dans un premier temps, nous avons souhaité nous assurer du bon fonctionnement de

l'algorithme d'estimation par maximum de vraisemblance implémenté. Pour cela, nous avons simulé 200 cohortes de 30425 travailleurs selon un modèle de Cox pour différentes valeurs du coefficient de risque β (0.0001, 0.0005, 0.001, 0.009) et des valeurs constantes pour les paramètres du taux de base : $\lambda_1 = 2.44 \times 10^{-7}$, $\lambda_2 = 2.48 \times 10^{-6}$, $\lambda_3 = 4.87 \times 10^{-5}$. Puis, nous avons ajusté ce même modèle de Cox aux données simulées. Les tables 5.2, 5.3, 5.4 et 5.5 indiquent le taux de couverture à 95% et le biais moyen (sur 200 jeux de données) obtenus pour chaque paramètre inconnu et ce, pour les 4 "vraies" valeurs considérées pour le coefficient de risque β

TABLE 5.2: Taux de couverture à 95% et biais moyen sur les paramètres d'un modèle de Cox pour une "vraie" valeur $\beta = 0.0001$

	Taux de couverture	**Biais moyen**
β	92%	2.48×10^{-4}
λ_1	96%	8.53×10^{-10}
λ_2	93%	7.82×10^{-10}
λ_3	96%	4.1×10^{-9}

TABLE 5.3: Taux de couverture à 95% et biais moyen sur les paramètres d'un modèle de Cox pour une "vraie" valeur $\beta = 0.0005$

	Taux de couverture	**Biais moyen**
β	93%	3.19×10^{-4}
λ_1	94%	-1.48×10^{-8}
λ_2	97%	-2.57×10^{-8}
λ_3	92%	-9.61×10^{-9}

TABLE 5.4: Taux de couverture à 95% et biais moyen (200 jeux de données) sur les paramètres d'un modèle de Cox pour une "vraie" valeur $\beta = 0.001$

	Taux de couverture	**Biais moyen**
β	91%	-4.02×10^{-5}
λ_1	93%	-1.61×10^{-7}
λ_2	94%	-5.03×10^{-7}
λ_3	94%	3.84×10^{-8}

TABLE 5.5: Taux de couverture à 95% et biais moyen (200 jeux de données) sur les paramètres d'un modèle de Cox pour une "vraie" valeur $\beta = 0.009$

	Taux de couverture	**Biais moyen**
β	90%	-1.67×10^{-5}
λ_1	95%	-2.43×10^{-7}
λ_2	93%	-5.72×10^{-7}
λ_3	96%	9.67×10^{-8}

A partir des tableaux obtenus, on remarque que les taux de couverture à 95% des paramètres sont tous supérieurs à 90% et proches du taux nominal de 95%. Cela indique

que la procédure d'estimation par maximum de vraisemblance semble "correcte" ainsi que l'algorithme de simulation : on retrouve bien les vraies valeurs de paramètres fixées pour la simulation dans environ 95% des cas (ce qui correspond à une erreur de première espèce $\alpha = 5\%$). Par ailleurs, les biais moyens observés sont petits et ne prennent pas un signe particulier.

On remarque que plus β est petit et plus le taux de couverture est grand. Plus β est proche de 0, plus le nombre de décès par cancer solide observé est petit. En effet, si $\beta \longrightarrow 0$ alors les risques instantanés de décès par cancer solide sont eux-mêmes proches de 0 car $\gamma_{ij} = \lambda_{l_j} \exp(\beta Z_{ij}) \longrightarrow \lambda_{l_j}$ avec λ_{l_j} proches de zéros. Or, plus le nombre de décès par cancer solide est petit et plus la variance d'estimation de β est élevée induisant un intervalle de confiance à 95% sur $\hat{\beta}$ "large". Par conséquent, la vraie valeur de β a plus de chance d'appartenir à cet intervalle d'où un taux de couverture élevé comparé à un intervalle de confiance "serré" relatif à une grande valeur de β. En suivant le même raisonnement que ci-dessus, plus le β choisi pour la simulation est petit et "moins bonne" est l'estimation du $\hat{\beta}$, car il y a moins de décès simulés. Ceci implique un biais plus grand pour les petites valeurs de β.

Dans un deuxième temps, nous avons simulé 200 jeux de données selon un modèle de Cox pour chacune des valeurs du coefficient de risque β définies dans la section 5.3.1 afin d'approcher asymptotiquement la puissance statistique de la cohorte EDF pour ces différentes valeurs, dans le cadre de la mise en évidence d'un risque de décès par cancer solide radio-induit associé à l'exposition aux rayonnements gamma.

TABLE 5.6: Puissance statistique estimée au niveau $\alpha = 0.05$ pour la cohorte EDF initiale i.e., P_{2003} (date de point : 2003) et la cohorte étendue i.e., P_{2014} (date de point : 2014) pour la mise en évidence de différentes valeurs $exp(\beta)$ de ratio (pour 1 mSv) de risques instantanés de décès par cancer solide radio-induit à partir d'un modèle de Cox.

$exp(\beta)$	1.0001	1.0005	1.0009	1.001	1.003	1.005	1.007	1.009	1.015
P_{2003}	8%	8%	15%	15%	56%	89%	99 %	100%	100%
P_{2014}	5.85%	10.22%	21.76%	23.2%	97%	100%	100%	100%	100%

Les valeurs de puissance statistique estimée au niveau $\alpha = 0.05$ pour quelques valeurs de β et dans le cas de l'utilisation d'un modèle de Cox sont indiquées dans la table 5.6 pour la cohorte initiale (date de point : 2003) et la cohorte étendue (date de point : 2014). Pour un ratio de risques instantanés de décès par cancer solide radio-induit de 1.0005

pour une dose de 1 milliSievert (mSv)- valeur à laquelle s'attendent les épidémiologistes du LEPID - la puissance statistique estimée (au niveau $\alpha = 0.05$) est de 8% si on considère une date de point à 2003 et de 10.22% si on considère une date de point à 2014. Cette puissance est très faible. Cela signifie que, même si un effet radio-induit existe, la probabilité pour que les données de la cohorte EDF permettent de conclure à l'existence de cet effet est très faible. Par ailleurs, dans le cas où les données disponibles ne permettraient pas de rejeter l'hypothèse H_0 : "Absence d'effet", la démarche statistique utilisée ne permet pas pour autant de conclure à l'absence d'effet. Enfin, une puissance statistique optimale de 80% serait obtenue dans le cadre de la mise en évidence d'un ratio de risques instantanés de 1.0046 pour 1 mSv pour la date de point de 2003 et, respectivement, de 1.0026 pour 1 mSv pour la date de point de 2014. Ces valeurs sont cependant trop élevées par rapport au ratio de risques instantanés attendu de 1.0005.

Les courbes de puissance statistique approchée obtenues ainsi que l'erreur de Monte-Carlo associée sont représentées dans :

— la figure 5.7 pour la cohorte initiale (date de point : 2003) ;

— la figure 5.8 pour la cohorte étendue (date de point : 2014) ;

— la figure 5.9 pour les cohortes initiales et étendues simultanément

Les courbes de puissance obtenues représentent la fiabilité de la procédure statistique choisie. Par conséquent, plus la convergence de la puissance est rapide et plus la procédure statistique (choix du modèle, choix de la statistique de test...) est fiable.

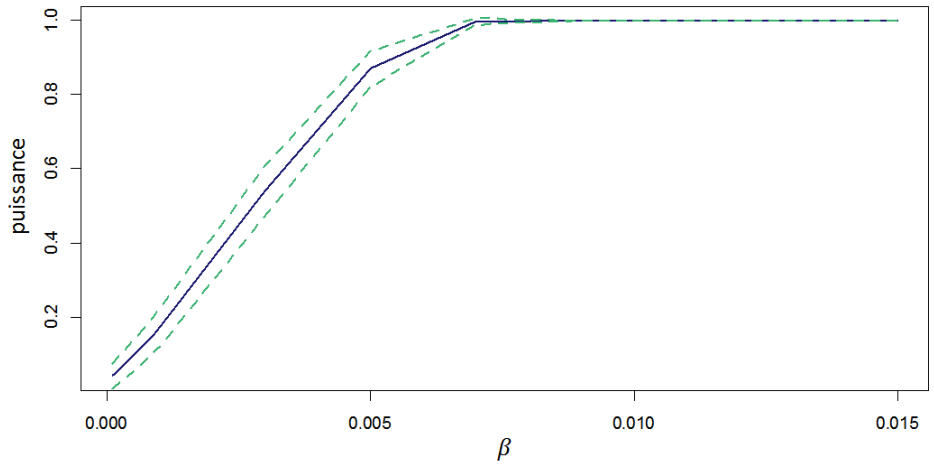

FIGURE 5.7: Courbe de puissance statistique estimée en fonction du coefficient de risque β (ligne continue bleue) et erreur de Monte-Carlo associée (pointillés verts) au niveau $\alpha = 0.05$ pour la cohorte EDF initiale (date de point : 2003) pour la mise en évidence d'un risque de décès par cancer solide radio-induit à partir d'un modèle de Cox

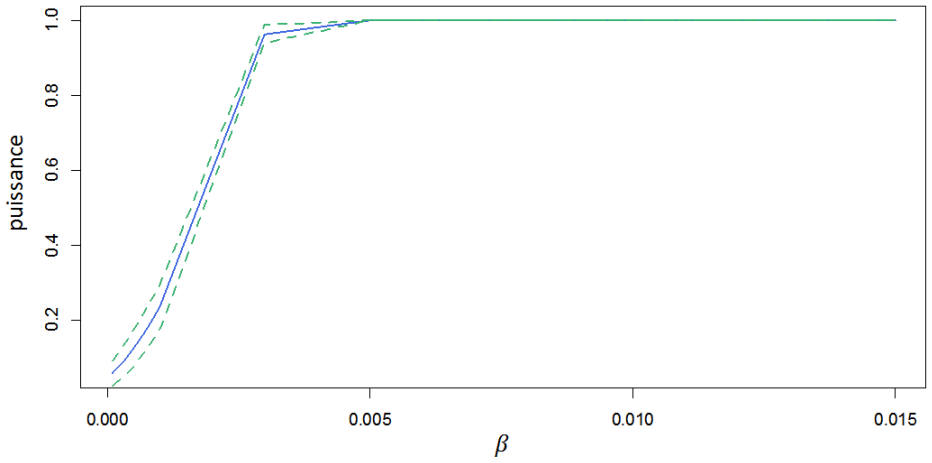

FIGURE 5.8: Courbe de puissance statistique estimée en fonction du coefficient de risque β (ligne continue bleue) et erreur de Monte-Carlo associée (pointillés verts) au niveau $\alpha = 0.05$ pour la cohorte EDF étendue (date de point : 2014) pour la mise en évidence d'un risque de décès par cancer solide radio-induit à partir d'un modèle de Cox

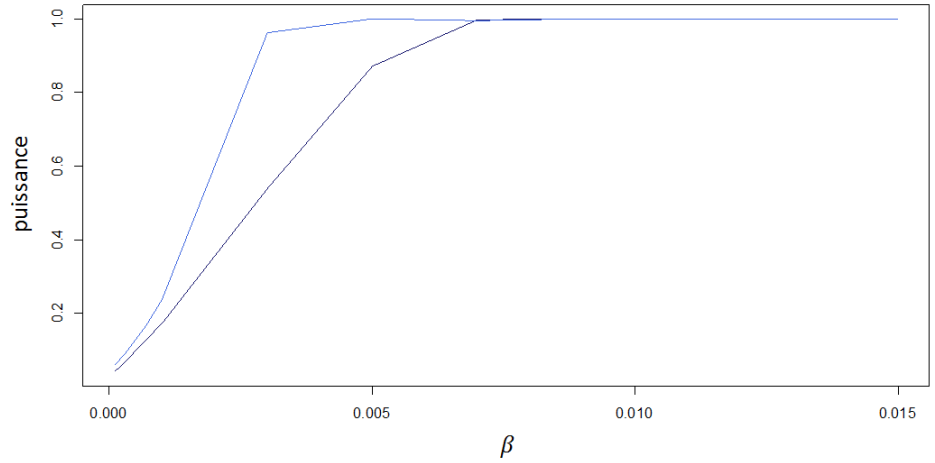

FIGURE 5.9: Comparaison des courbes de puissance statistique estimées en fonction du coefficient de risque β pour la cohorte EDF initiale (en noir) et la cohorte étendue (en bleue) pour la mise en évidence d'un risque de décès par cancer solide radio-induit à partir d'un modèle de Cox

On retrouve un résultat cohérent : la puissance statistique approchée de la cohorte EDF est croissante avec la valeur du coefficient de risque β. Cela s'explique par le fait que plus le coefficient de risque β est "grand", plus le nombre de décès par cancer solide est élevé et donc "meilleure" est l'estimation du paramètre β (avec un écart-type d'estimation $\hat{\sigma}_i$ petit). Ainsi, une "meilleure" estimation induit un intervalle de confiance $]\hat{\beta}_i - z_{\frac{\alpha}{2}} \times \hat{\sigma}_i, \hat{\beta}_i + z_{\frac{\alpha}{2}} \times \hat{\sigma}_i[$ plus petit. 0 a donc moins de chance d'appartenir à cet intervalle de confiance et par conséquent la puissance est "grande" pour une valeur élevée de β.

Enfin, comme on pouvait s'y attendre, la puissance augmente avec le temps de suivi des travailleurs : c'est ce qu'on retrouve dans la figure 5.9. En effet, la courbe de puissance à la date de point 2014 est supérieure en tout point à la courbe de puissance à la date de point en 2003. En effet, en prolongeant le temps de suivi des travailleurs, on a plus de chance d'observer des décès par cancer solide. On peut invoquer deux raisons principales à cela. Tout d'abord, même en absence d'exposition additionnelle aux rayonnements gamma, le risque instantané de décès par cancer solide croit avec l'âge comme défini à travers le risque instantané de base de décès par cancer solide, constant par morceaux (cf. figure 5.2). Par ailleurs, le risque instantané de décès par cancer solide $\gamma_{ij} = \lambda_{l_j} \exp(\beta Z_{ij})$ croît avec la dose cumulée d'exposition aux rayonnements gamma Z_{ij}. Or, celle-ci a des chances d'augmenter quand le temps de suivi augmente, pour les travailleurs encore en activité chez EDF. Par conséquent, plus le temps de suivi augmente, meilleure est l'estimation du coefficient de risque β et ainsi, la puissance statistique augmente. Des résultats de puissance supplémentaires (pour le modèle en EHR à la date de point en 2003) sont présentés en annexe (cf. annexe A.7).

5.4 Discussion

Les calculs de puissance statistique réalisés dans ce stage dépendent des modèles de survie considéré et du niveau α choisi. Deux principales limites peuvent être évoquées quant aux hypothèses de modélisation adoptées.

La première limite concerne le choix d'une fonction de risque instantané de base constante par morceaux sur les intervalles d'âge (en années) suivants : $]0, 40]$,$]40, 60]$ et $]60, +\infty[$. Le choix d'une fonction constante par morceaux permet de faciliter le calcul intégral des fonctions de survie des modèles de Cox ou en EHR. Néanmoins, il ne permet de considérer qu'une partition "grossière" du temps dans le cas de l'estimation d'un risque faible de décès. Le choix d'une partition plus fine du temps conduit à des difficultés d'inférence auxquelles nous avons été confrontées pendant ce stage. En effet, notre choix de modélisation initial était de supposer un risque instantané de base h_0 constant par morceaux sur les quatre intervalles de temps suivants : $]0, 40],]40, 60],]60, 85]$ et $]85, +\infty[$. Cependant, comme aucun voire très peu de travailleurs EDF sont décédés par cancer solide après 85 ans car souvent plus jeunes que 85 ans à la date de point de l'étude (exemple : l'âge moyen à la date de point de 2003 est

de 45.78 ans), le paramètre λ_4 définissant le taux de base après 85 ans ne s'estime pas correctement. Nous avons par ailleurs pu constater que si l'un des paramètres du taux de base est mal estimé alors les autres paramètres s'estiment mal également. La Figure 5.10 représente les profils de vraisemblance pour différentes valeurs des paramètres β, λ_1, λ_2, λ_3 et λ_4 et obtenus à partir d'un jeu de données de survie simulées selon le modèle de Cox pour 30425 travailleurs et des "vraies" valeurs de paramètres fixées à $\beta = 0.01$, $\lambda_1 = 2.44 \times 10^{-7}$, $\lambda_2 = 2.48 \times 10^{-6}$, $\lambda_3 = 1.57 \times 10^{-5}$, $\lambda_4 = 8.16 \times 10^{-5}$. On remarque que a) le maximum de vraisemblance ne coincide pas toujours avec la "vraie" valeur des paramètres (cas des paramètres λ_3 et λ_4) ; b) que le profil de vraisemblance pour un paramètre d'intérêt θ (i.e., fonction de vraisemblance en θ optimisée par rapport aux autres paramètres) est parfois plate (cas du paramètre β).

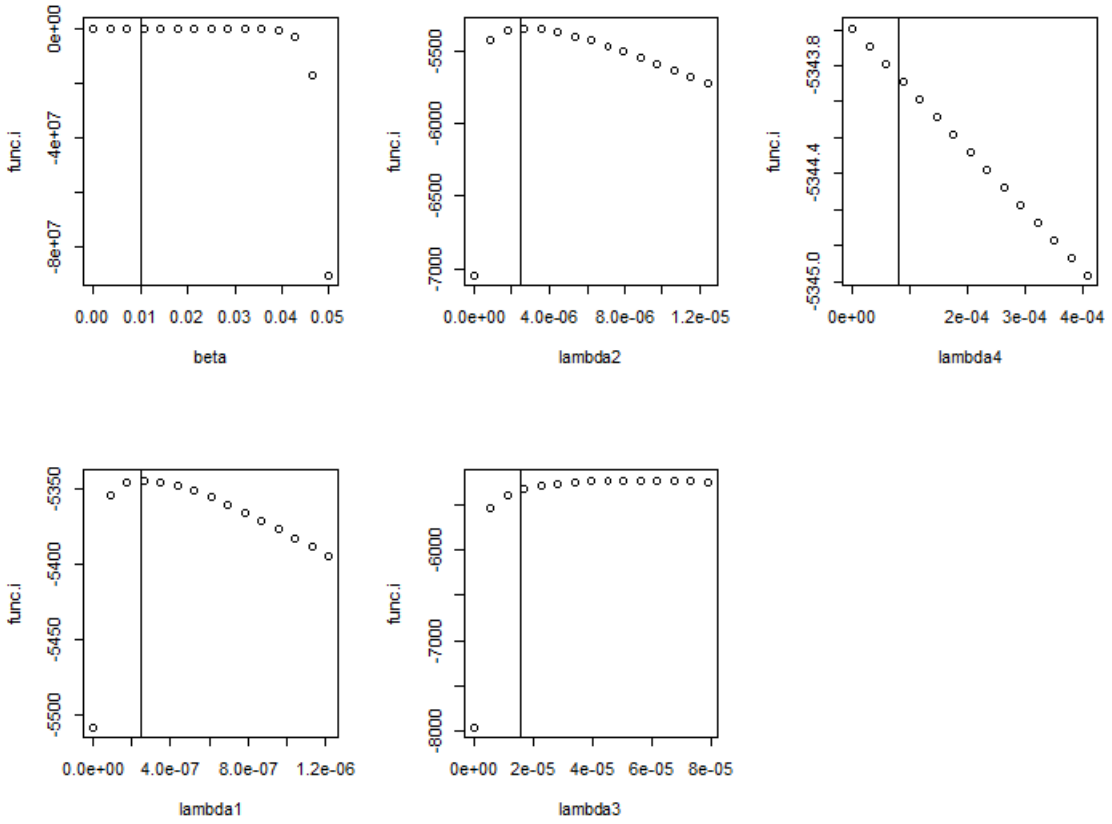

FIGURE 5.10: Profils de vraisemblance au voisinage des "vraies" valeurs de paramètre $\beta = 0.01$, $\lambda_1 = 2.44 \times 10^{-7}$, $\lambda_2 = 2.48 \times 10^{-6}$, $\lambda_3 = 1.57 \times 10^{-5}$, $\lambda_4 = 8.16 \times 10^{-5}$ pour des données de survie simulées selon le modèle de Cox

Par ailleurs, dans le cas de deux (respectivement un) décès observés après 85 ans, la vraisemblance en fonction de λ_4 est une droite (respectivement constante) et donc sa dérivée seconde est égale à 0. On se retrouve alors avec un système singulier . On ne peut

donc pas inverser la matrice de Fisher observée (i.e. la hessienne) et donc pas estimer l'écart-type d'estimation des paramètres. C'est la raison pour laquelle nous avons finalement choisi un taux de base constant par morceaux sur trois intervalles : $]0, 40]$,$]40, 60]$ et $]60, +\infty[$. Néanmoins, nous avons vu que la distribution des données de décès par cancer solide simulées est très influencée par cette forme de taux de base et qu'elle ressemble finalement peu à celle des données de la cohorte EDF (cf. Figure 3.2) . Compte-tenu de ces remarques, il semblerait judicieux de choisir une autre forme de fonction pour le risque instantané de base. Une amélioration possible serait de choisir une fonction de risque paramétrique qui soit toujours croissante mais continue dans le temps ce qui permettrait d'estimer par extrapolation le risque de base pour les âges au décès supérieurs à 85 ans même en l'absence de décès observés. Cette fonction devra être nulle en zéro et son inverse différentiable afin de pouvoir appliquer la méthode de simulation proposée. Enfin, elle devra également satisfaire à certaines exigences pour que l'algorithme acceptation rejet utilisé pour tronquer les âges au décès simulés puisse converger rapidement.

La deuxième limite concerne le modèle choisi pour décrire la censure aléatoire. Nous avons choisi une loi uniforme. Néanmoins, cette hypothèse peut être critiquée car on s'attend plutôt à une augmentation du risque instantané de censure aléatoire avec le temps puisque cette censure est notamment liée au risque de décéder par d'autres causes que les cancers solides. Une idée pour améliorer ceci serait de simuler la censure aléatoire selon une loi exponentielle par morceaux avec intensités croissantes avec le temps ou en considérant toute fonction de risque instantané continue et croissante permettant de simuler des temps de censure plausibles (comme pour l'âge au décès).

En principe, un calcul de puissance s'effectue en amont de l'acquisition de données. Pour ce stage, on disposait déjà des données complètes pour la cohorte EDF initiale (i.e., date de point 2003) et des données d'exposition aux rayonnements gamma pour la cohorte étendue (i.e., date de point 2014). On a donc eu la chance de pouvoir utiliser les valeurs de doses des bases de données disponibles. Une simulation de doses aurait du être effectuée si celles-ci n'avaient pas été disponibles.

Dans le cadre d'un calcul de puissance statistique pour la mise en évidence d'un effet sanitaire radio-induit, on a choisi la statistique de test de Wald. Néanmoins, dans le cadre d'un test avec une hypothèse alternative simple et selon le théorème de

Neyman-Pearson, le test optimal est celui du rapport de vraisemblance [28]. Mais dans le cas d'un modèle de Cox ou en EHR, on ne peut isoler une statistique de test à partir du rapport de vraisemblance et par conséquent, on ne peut ni calculer la région de rejet, ni la puissance. C'est la raison pour laquelle nous avons choisi la statistique de Wald, car celle-ci a de bonnes propriétés et nous permet d'approximer la puissance. Une alternative à la statistique de Wald est la statistique de score normalisée, fréquemment utilisée pour le calcul de puissance dans le cadre des études épidémiologiques [19]. Cependant la statistique de test de Wald nous permet d'obtenir une approximation beaucoup plus simple de la puissance. Une comparaison des résultats de puissance obtenus à partir de statistiques de test différentes aurait été intéressante à effectuer.

La puissance est la probabilité de rejeter H_0 sachant qu'on est sous l'hypothèse alternative H_1. Cela revient, dans le cadre d'une hypothèse alternative simple, à rejeter H_0 sachant que β est égale à une valeur β_1 non nulle fixée. Or, le fait de rejeter H_0 dépend des données (car dépend de la statistique de test) donc calculer la puissance revient à calculer la probabilité d'une fonction des données sachant β_1 fixé. Ceci présente deux inconvénients majeurs :

— Le premier est le fait qu'on raisonne à $\beta = \beta_1$ fixé, alors que β est inconnu.
— Le deuxième est qu'on calcule la probabilité des données , qui sont encore inconnues dans un contexte de calcul de puissance, en fixant la valeur des paramètres inconnus.

On peut donc se poser la question suivante : Etant donné qu'on veut savoir s'il existe un risque de décès radio-induit lors d'une étude de cohorte professionnelle, ne serait-il pas plus naturel de se placer dans le cadre bayésien en calculant la probabilité pour que $\beta = 0$ sachant les données et en spécifiant une loi *à priori* sur β ? Dans ce contexte, un équivalent bayésien de la puissance statistique serait la probabilité de $\beta \neq 0$ sachant les données observées. Se placer dans le cadre bayésien, permettrait de s'affranchir des deux inconvénients ci-dessus. En effet, β serait aléatoire (donc on ne le fixera pas) et on calculera la probabilité de β une fois les données observées au lieu de la probabilité des données (qui sont inconnues dans le cas d'un calcul de puissance) sachant β. Cette procédure a ainsi un prix : celui du choix de la loi *a priori* sur β !

A noter enfin que pour le premier inconvénient, des approches semi-bayésiennes existent pour intégrer la puissance statistique dans le cadre fréquentiste sur plusieurs valeurs de β [7]

Chapitre 6

Calibration optimale d'un protocole d'étude de cohorte

Le but de ce chapitre est de décrire les premiers éléments de formalisation mathématique concernant la question (b) introduite dans la partie [2.3], c'est-à-dire trouver le nombre de sujets et le temps de suivi à ajouter à la cohorte EDF pour obtenir une puissance optimale pour la mise en évidence d'un risque sanitaire radio- induit.

6.1 Choix d'une fonction d'utilité et d'un critère à maximiser

Dans le cadre de la théorie de la décision fréquentiste appliquée au problème de calibration optimale de protocole d'étude de cohorte abordé dans ce stage, le décideur recherche le protocole d'étude qui permet de maximiser une fonction d'utilité u en moyenne (i.e., intégrée sur toutes les configurations de données possibles y) pour une valeur de paramètres θ. Pour cela, il utilise la fonction d'utilité (ou fonction de gain) classique définie par :

$$U(\eta, \theta) = \int_{\mathcal{Y}} u(\eta, \theta, y)[y|\theta, \eta]dy$$

où η désigne un résumé du protocole d'étude (par exemple : nombre de sujets et temps de suivi à ajouter à la cohorte EDF), $[y|\theta, \eta]$ désigne la vraisemblance des données y observées (par exemple : les temps de survie et indicateurs de non-censure) sachant un

protocole d'étude η et un vecteur de paramètres θ (par exemple : $\theta = (\beta, \lambda_1, \lambda_2, \lambda_3)$). u est une fonction d'utilité choisie et $\mathcal{Y}$ l'espace des réalisations des variables observables y.

On peut remarquer que l'utilité classique définie plus haut nous permet uniquement de déterminer un protocole d'étude η localement optimal, dans le sens où l'utilité apportée par ce dernier est uniquement évaluée pour une certaine valeur fixée des paramètres inconnus θ.

Si l'objectif est de trouver le protocole d'études η optimal pour la mise en évidence d'un risque sanitaire radio-induit, on choisira u de telle sorte que maximiser U dans l'espace des protocoles d'étude possibles η nous permette de maximiser la puissance statistique d'un test pour la mise en évidence, s'il existe, d'un risque sanitaire radio-induit. Or, comme vu dans la partie [5.1.2], estimer "au mieux" les paramètres d'un modèle de Cox ou en EHR permet d'augmenter la puissance statistique d'intérêt. Une solution possible pour estimer "au mieux" ces paramètres est de chercher à minimiser leur écart-type d'estimation respectif. Dans ce contexte, une fonction d'utilité u possible est définie comme la perte quadratique engendrée par l'estimateur ponctuel $\hat{\theta}$ soit l'opposé de l'erreur quadratique d'estimation :

$$u(\eta, \theta, y) = -(\theta - \hat{\theta})^t(\theta - \hat{\theta})$$

où $\hat{\theta}$ est un estimateur de θ.

Soit $\hat{\theta}$ l'estimateur du maximum de vraisemblance du vecteur de paramètres θ (supposé de dimension p). Dans le cas d'un modèle probabiliste régulier, on sait que $\hat{\theta}$ suit asymptotiquement la loi Gaussienne suivante : $\mathcal{N}_p(\theta, I(\hat{\theta}, \eta)^{-1})$ avec $I(\hat{\theta}, \eta)$ la matrice d'information de Fisher observée associée au modèle d'observations supposé pour les variables observables Y. Par ailleurs, à y et η fixés, un estimateur ponctuel fréquentiste $\hat{\theta}$ permettant de minimiser asymptotiquement la perte quadratique est l'estimateur du maximum de vraisemblance.

Supposons que la fonction d'utilité est $u(\eta, \theta, y) = -(\theta - \hat{\theta})^t(\theta - \hat{\theta})$. Dans ce cas, l'utilité classique associée est définie par :

$$\begin{aligned}
U(\eta, \theta) &= \int_{\mathcal{Y}} u(\eta, \theta, y)[y|\theta, \eta]dy \\
&= -\int_{\mathcal{Y}} (\theta - \hat{\theta})^t(\theta - \hat{\theta})[y|\theta, \eta]dy \\
&= -\mathbb{E}_{Y|\theta}\Big[\sum_{i=1}^{p}\sum_{j=1}^{p}[(\theta_i - \hat{\theta}_i)(\theta_j - \hat{\theta}_j)]\Big] \\
&= -\sum_{i=1}^{p}\sum_{j\neq i}\mathbb{E}_{Y|\theta}\big[(\theta_i - \hat{\theta}_i)(\theta_j - \hat{\theta}_j)\big] - \sum_{i=1}^{p}\mathbb{E}_{Y|\theta}\big[(\theta_i - \hat{\theta}_i)^2\big] \\
&\approx -\sum_{i=1}^{p}\mathbb{E}_{Y|\theta}\big[\mathbb{V}[(\hat{\theta}_i)]\big] \\
&\approx -\mathbb{E}_{Y|\theta}\big[tr\big(I(\hat{\theta}, \eta)^{-1}\big)\big] \\
&\approx -tr\Big[\frac{1}{\mathbb{E}_{Y|\theta}\big(I(\hat{\theta}, \eta)\big)}\Big]
\end{aligned}$$

Remarque : $\sum_{i=1}^{p}\sum_{j\neq i}\mathbb{E}_{Y|\theta}\big[(\theta_i - \hat{\theta}_i)(\theta_j - \hat{\theta}_j)\big] \approx 0$ car les $\hat{\theta}_i$ sont indépendants et $\mathbb{E}_Y(\hat{\theta}_i) \approx \theta_i$ pour tout i.

Maximiser $U(\eta, \theta)$ en η revient donc à maximiser le critère dit de A-optimalité local suivant :

$$\phi(\eta, \theta) = -tr\Big[\frac{1}{\mathbb{E}_{Y|\theta}\big(I(\hat{\theta}, \eta)\big)}\Big]$$

6.2 Quelle contrainte pour la recherche d'un protocole d'étude optimal ?

Considérons l'ensemble des couples $\eta = (\Delta n, \Delta t) \in \{1, \ldots, B_1\} \times \{1, \ldots, B_2\}$ de nombres d'individus et de nombres d'années de suivi qu'il serait possible d'ajouter à la cohorte EDF initiale. B_1 désigne, par exemple, le nombre total de travailleurs du nucléaire dans le monde et B_2 le nombre maximal d'années de suivi à ajouter pour que ces B_1 travailleurs décèdent.

En considérant uniquement la fonction d'utilité u définie plus haut, on s'attend intuitivement à ce que le critère de A-optimalité soit maximal pour le couple

$(\Delta n, \Delta t) = (B_1, B_2)$. En effet, plus on augmente le nombre d'individus et le temps de suivi, plus le nombre de décès observés par cancer solide augmente et donc plus l'écart-type d'estimation de θ diminue et la puissance statistique augmente. Or, en pratique, on ne peut atteindre ce maximum pour des raisons de contraintes de temps et de contraintes financières. Dans le cadre de ce stage, il a donc été nécessaire de définir une contrainte opérationnelle sur le couple $(\Delta n, \Delta t)$ pour la recherche d'un protocole d'étude optimal pour la mise en évidence, s'il existe, d'un risque sanitaire radio-induit.

Pour une première approche du problème, nous avons considéré la contrainte suivante : supposons qu'on puisse faire 100 nouveaux points de suivi (par exemple au premier janvier de chaque année), qu'on puisse mesurer la dose de rayonnements gamma en ces points et que l'indicateur de non-censure de chaque individu soit disponible en ces points. Naturellement, on peut se poser la question suivante : Parmi toutes les situations possibles, quelle est celle qui permet de maximiser la puissance de l'étude pour la mise en évidence, s'il existe, d'un risque sanitaire radio-induit ? Par exemple, vaut-il mieux suivre :

- un individu pendant cent ans ? ;
- cinq individus pendant vingt ans ? ;
- dix individus pendant dix ans ?
- cent individus pendant un an ?

Dans ce contexte, c'est la probabilité de censure aléatoire qui jouera le rôle de contrainte "naturelle". En effet, intuitivement, il semble inutile de suivre cent individus pendant une année supplémentaire puisqu'on risque alors d'observer aucun décès supplémentaire. De même, il est inutile d'observer cent fois un même individu car cela ne permettra d'ajouter à coup sûr qu'un seul décès supplémentaire. L'ensemble des couples $\eta = (\Delta n, \Delta t)$ à considérer se résume donc à l'ensemble des couples $(\Delta n, \frac{K}{\Delta n})$ avec K un entier naturel fixé (i.e., nombre de points de suivi supplémentaires) et $\Delta n \in \mathcal{D}_K$ avec $\mathcal{D}_K$ l'ensemble des diviseurs possibles de K. Cette contrainte permet de simplifier l'espace des protocoles d'étude η sur lequel doit être optimisé le critère de A-optimalité ϕ. En effet, celui-ci se résume alors à l'ensemble des diviseurs possibles de K. On se ramène donc à une optimisation discrète de ϕ sur l'ensemble $\mathcal{D}_K$.

6.3 Application à la cohorte EDF

Les résultats ci-dessous ont été obtenus en se fixant le choix de $K = 1100$ points de suivi supplémentaires. Comme nous disposons de valeurs de doses de rayonnements gamma entre 2003 et 2014, nous nous sommes fixées, pour cette première application numérique, une valeur maximale pour $\Delta t = 11$ ce qui revient à considérer uniquement les valeurs de Δn suivantes : $\{100, 110, 220, 275, 550, 1100\}$

Les travailleurs ajoutés à la cohorte initiale ont été choisis aléatoirement parmi les individus non exposés à la date de point de 2003 (car les nouveaux individus inclus dans la cohorte sont non exposés ou sont tels que leur exposition est inconnue).

Nous avons recherché le protocole d'étude optimal pour la mise en évidence, si elle existe, d'une association entre l'exposition aux rayonnements gamma et le risque de décès par cancer solide chez les agents statutaires EDF, dans le contexte spécifique où le coefficient de risque $\beta = 0.0005$ (c'est la valeur à laquelle les épidemiologistes s'attendent) et pour des valeurs de paramètres fixées pour le risque instantané de base (constant par morceaux) : $\lambda_1 = 2.44 \times 10^{-7}$, $\lambda_2 = 2.48 \times 10^{-6}$, $\lambda_3 = 4.87 \times 10^{-5}$. 100 cohortes EDF "étendues" ont été simulées selon un modèle de Cox tronqué à gauche et avec covariables dépendantes du temps, pour chaque couple de valeurs $\eta = (\Delta n, \Delta t)$ considérées. En effet, le calcul de l'utilité classique nécessite, pour chaque couple η, d'intégrer u sur l'ensemble des réalisations de données de survie possibles à $\theta = (\beta, \lambda_1, \lambda_2, \lambda_3)$ fixé.

La trace de l'inverse de la matrice d'information de Fisher, dont l'expression mathématique en fonction de Δn et Δt a été définie dans l'Annexe A.6.1), a été calculée pour chaque cohorte "étendue" simulée. Puis le critère de A-optimalité a été approximé, pour chaque couple η, en moyennant les 100 valeurs de traces obtenues.

Les 100 simulations selon un modèle de Cox pour $k = 1100$ points de suivi ont été effetuées pour chaque Δn, mais les résultas n'ont pas pu être obtenus à cause du temps de calcul long des traces de matrices de Fisher. Ces résultats seront cependant disponibles pour la soutenance.

6.4 Discussion

Le choix d'une contrainte d'optimalité sur le problème posé a été l'une des difficultés majeures rencontrées dans cette partie. En effet, l'élicitation d'une contrainte budgétaire pertinente n'a pas été possible dans le cadre de ce stage ce qui a nécessité de contraindre le problème d'optimisation de protocole d'étude de façon *a priori* moins opérationnelle.

La contrainte d'optimalité considérée a conduit à proposer une version discrète du problème de calibration optimale de protocole d'étude de cohorte. Une optimisation discrète est néanmoins difficile en pratique à cause de l'absence du gradient et nécessite parfois l'utilisation d'algorithmes compliqués en grande dimension [22]. Dans notre cas, l'optimisation a consisté à calculer l'inverse de la matrice d'information de Fisher pour seulement un petit nombre de valeurs de Δn puis de choisir le Δn qui maximise le critère de A-optimalité approché. Dans un autre contexte, un moyen de contourner les difficultés engendrées par une optimisation discrète est de rendre le problème continu, par exemple en affectant un poids $\eta_i \in [0, 1]$ à chaque situation possible [35], ou en optimisant directement sur un ensemble de mesures en utilisant le théorème d'équivalence de Whittle (1973) [8].

Le choix d'une fonction de coût quadratique nous a permis d'obtenir une approximation asympotique directe de la fonction d'utilité classique associée. Ce choix peut cependant être critiqué et d'autres fonctions d'utilité pourraient sans doute être utilisées, par exemple :

- Utilité de Shannon (dans le cadre bayésien, elle correspond au design bayésien D-optimal) : Le plan D-optimal correspond au déterminant de la matrice d'information de Fisher [1] ;
- Plan d'expérience bayésien C-optimale [35] : analogue à la fonction d'utilité de Shannon, sauf qu'on intègre sur la distribution prédictive au lieu d'intégrer sur le posterior ;
- Utilité proposée par Verdinelli et Kadane [34]

$$U(\eta) = \int [\rho y^t \mathbf{1} + \beta log p(\theta|y, \eta)] p(y, \theta|\eta) dy d\theta$$

où ρ est un poids et β représente la contribution relative que l'experimentateur voudrait attribué aux deux composantes de U.

L'utilité classique utilisée dans le cadre fréquentiste permet uniquement d'obtenir un protocole d'étude optimal localement, dans le sens où on choisit le maximum de l'utilité classique pour un vecteur de paramètres θ fixé. La démarche fréquentiste peut donc être critiquée. Une alternative prometteuse serait de se placer dans le cadre bayésien afin d'optimiser une utilité dite 'espérée' obtenue en intégrant l'utilité classique sur la loi a *priori* de θ afin d'obtenir un maximum global [1].

Chapitre 7

Conclusion

J'ai rencontré deux difficultés techniques majeures au cours de ce stage :

- l'implémentation en R d'une méthode adaptée pour la simulation puis l'estimation fréquentiste de données de survie **tronquées à gauche** et avec covariables **dépendantes du temps** a pris beaucoup plus de temps que prévu dans le cadre de ce stage. Très peu d'articles expliquent la simulation de données de survie avec covariables continues et dépendantes du temps et dans la plupart des cas, ces covariables sont binaires [36]).
- L'inférence fréquentiste d'un modèle de survie à covariables dépendantes du temps a été laborieuse. En effet, avant de se résoudre à faire l'inférence en utilisant la vraisemblance, j'ai essayé d'utiliser la fonction *coxph* de R pour l'estimation du paramètre du modèle β. Cependant, dans le cas d'un modèle de Cox à covariables dépendantes du temps la gestion de la base de données était très compliquée, de plus, la fonction *coxph* ne fournissait pas les bons taux de couverture.

Malgré ces difficultés, je suis parvenue à apporter de premiers éléments de réponse à deux questions d'intérêt actuelles pour le LEPID, à savoir : a) quelle est la puissance statistique de la cohorte EDF pour la mise en évidence, si elle existe, d'une association entre le risque de décès par cancer solide et d'une exposition chronique et à faibles doses aux rayonnements gamma ? et b) comment améliorer, de façon optimale, le protocole de l'étude de cohorte EDF afin d'estimer au mieux le risque de décès par cancer solide associé à l'exposition aux rayonnements gamma ? Ainsi, j'ai obtenu les premiers résultats de puissance statistique pour la cohorte EDF à la date de point 2003 et à la date de point 2014 (extension de suivi en cours par le LEPID). Pour un ratio de risques

instantanés de décès par cancer solide radio-induit de 1.0005 pour une dose de 1 milliSievert - valeur représentative des ordres de grandeurs auxquels s'attendent les épidémiologistes compte tenu de la littérature scientifique internationale - la puissance statistique associée est de 8% si on considère une date de point à 2003 et de 10.22% si on considère une date de point à 2014. Cette puissance est très faible. Cela signifie que, même si un effet radio-induit existe, la probabilité pour que les données de la cohorte EDF seule permettent de conclure à l'existence de cet effet est très faible. Enfin, j'ai proposé une première formalisation mathématique du problème de calibration optimale d'un protocole d'étude de cohorte professionnelle pour la mise en évidence d'un risque sanitaire radio-induit. Enfin, ce stage a conduit à proposer et implémenter une méthode permettant de simuler des données de survie tronquées à gauche selon un modèle de Cox ou en excès de risque instantané avec covariables dépendantes du temps tels qu'utilisés en épidémiologie des rayonnements ionisants.

Ce stage m'a permis d'améliorer mes compétences en gestion de bases de données et en programmation sous R. J'ai notamment réussi à coder plusieurs formules très compliquées (matrice de Fisher du modèle de cox par exemple), ce que j'étais incapable de faire avant le début de mon stage. De plus, j'ai pu améliorer mes aptitudes en recherches bibliographiques, et j'ai approfondi mes connaissances dans le domaine de l'épidémiologie et ce, notamment, concernant les risques de cancers et les risques sanitaires radio-induits.

En ce qui concerne la statistique, j'ai pu appliquer ce que j'ai appris dans mon cours de survie à l'Université de Sherbrooke et approfondir mes connaissances dans ce domaine.

L'une de mes motivations dans le choix de ce stage était d'approfondir mes compétences en statistique bayésienne. Or, faute de temps, cela n'a pas pu être accompli.

Annexe A

Annexe

A.1 Quelques généralités sur les modèles de survie

Le terme de durée de survie désigne le temps écoulé jusqu'à la survenue d'un événement précis. L'événement étudié (communément appelé "décès") est le passage irréversible entre deux états (communément nommé "vivant" et "décès"). Par la suite, et comme cela sera le cas dans le cas d'étude EDF, on parlera de "décès". Néanmoins, l'événement terminal n'est pas forcément la mort : il peut s'agir de l'apparition d'une maladie (par exemple, le temps avant une rechute ou un rejet de greffe), d'une guérison (temps entre le diagnostic et la guérison), de la panne d'une machine (durée de fonctionnement d'une machine, en fiabilité) ou de la survenue d'un sinistre (temps entre deux sinistres, en actuariat).

L'analyse des données (durées) de survie est l'étude du délai de la survenue de cet événement. Dans le domaine biomédical, on étudie ces durées dans le contexte des études longitudinales comme les enquêtes de cohorte (suivi de patients dans le temps) ou les essais thérapeutiques (tester l'efficacité d'un médicament).

Quelques définitions sont couramment utilisées dans les études de survie.

— **Date d'origine** : elle correspond à l'origine de l'échelle de temps considérée. Elle peut être la date de naissance, le début d'une exposition à un facteur de risque, la date d'une opération chirurgicale, la date de début d'une maladie ou la date d'entrée dans l'étude. Chaque individu peut donc avoir une date d'origine différente (pas important car c'est la durée qui nous intéresse).

— **Date de point** : c'est la date de fin d'étude, c'est-à-dire au-delà de laquelle on arrêtera le suivi des individus de la cohorte.
— **Date des dernières nouvelles** : c'est la date la plus récente à laquelle des informations sur un sujet ont été recueillies.

Les temps de décès observés à partir d'une origine appropriée ont deux caractéristiques :

— La première est qu'ils sont non négatifs et tels qu'une hypothèse de normalité n'est généralement pas raisonnable en raison d'une asymétrie prononcée.
— La seconde est structurelle et tient au fait que, pour certains individus, le décès ne se produit pas pendant la période d'observation et en conséquence certaines données sont censurées.

A.1.1 Censure

Une des caractéristiques des données de survie est l'existence d'observations incomplètes. En effet, les données sont souvent recueillies partiellement, notamment, à cause de processus de censure. Les données censurées proviennent du fait qu'on n'a pas accès à toute l'information : au lieu d'observer un temps de décès X, on observe une borne minimale ou maximale pour ce temps.

La censure est le phénomène le plus couramment rencontré lors du recueil de données de survie. Pour l'individu i, considérons

— son temps de décès X_i
— son temps de censure C_i
— son temps de survie $T_i = min(X_i, C_i)$

Définition : Censure à droite
Un temps de décès X_i est dit censuré à droite si l'individu i n'est pas décédé à sa dernière date d'observation (ex. fin de l'étude ...).

En présence de censure à droite, les temps de décès ne sont pas tous observés ; pour certains d'entre eux, on sait seulement qu'ils sont supérieurs à une certaine valeur connue.

Considérons une étude relative à la durée de survie de patients soumis à un traitement particulier. L'évènement d'intérêt est la mort de la personne. Tous les individus sont suivis pendant les 80 semaines suivant la première administration du traitement. On considère plus particulièrement 3 sujets qui vont permettre d'illustrer certaines des

caractéristiques les plus fréquentes des données de survie et notamment deux cas possibles de censure à droite.

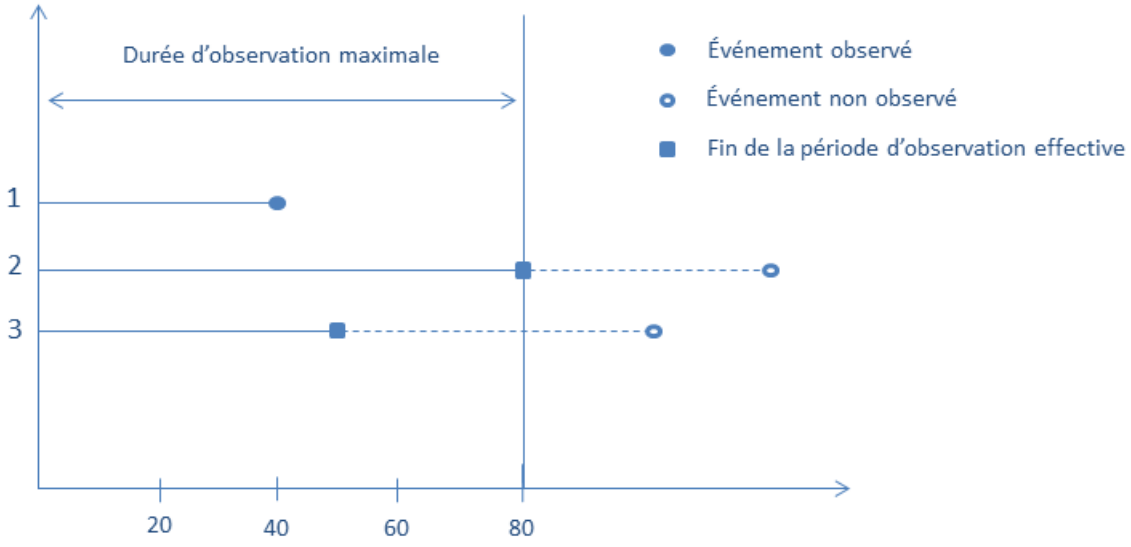

FIGURE A.1: Illustration de différents types de censure

— L'individu 1 est décédé 40 semaines après le début du traitement. Il s'agit d'une observation non censurée ($X_1 = 40$).

— La deuxième personne est toujours vivante au terme des 80 semaines d'observation. Elle décédera après 120 semaines mais cette information n'est pas connue à la date de point de l'étude. Même incomplète, l'information est utile puisque l'on sait que le temps de décès est supérieur à 80 semaines. Il faut donc tenir compte de cette information sous peine par exemple de biaiser vers le bas l'estimation du temps de décès moyen de tous les individus de la cohorte. Il s'agit d'une censure déterministe ou censure de type I, car elle ne dépend pas de l'individu considéré mais de la date de point de l'étude fixée par l'épidémiologiste ($X_2 = 120 > C = 80$).

— La troisième personne décède après 100 semaines mais cet évènement n'est pas enregistré dans la base de données car le patient concerné n'a pu être effectivement suivi que pendant 50 semaines, perdu de vue au delà de cette date. C'est un exemple de censure aléatoire ou censure de type III car elle échappe au contrôle de l'épidémiologiste. Là encore l'information est incomplète mais non nulle, car par exemple, savoir que cet individu a survécu au moins 50 semaines est pertinent pour l'estimation du taux de décès à 40 semaines ($X_3 = 100 > C_3 = 50$).

Comme la censure à droite, on peut aussi définir la censure à gauche et la censure par intervalle.

Dans le cadre de ce stage, les travailleurs de la cohorte EDF entrent dans l'étude à un certain âge et sont suivis jusqu'à l'âge minimal entre leur décès, la date de point de l'étude ou tout autre cause de censure. Si le travailleur ne décède pas de la cause d'intérêt (ici, cancer solide) pendant l'étude, alors il sera censuré à droite. Cette censure peut prendre deux formes : une censure déterministe s'il décède après la date de point et une censure aléatoire si, pendant l'étude, il décède d'une cause différente de la cause d'intérêt ou s'il est perdu de vue (ex., changement d'entreprise).

A.1.2 Troncature

Les troncatures diffèrent des censures au sens où elles concernent l'échantillonnage lui même. Ainsi, une variable X est tronquée par un sous ensemble éventuellement aléatoire A si au lieu de X, on observe X uniquement si $X \in A$. Les points de l'échantillon "tronqué" appartiennent donc tous à A.

Il ne faut pas confondre censure et troncature. S'il y a troncature, alors on sait avec exactitude que $\mathbb{P}(X \notin A) = 0$. Alors que s'il y a censure, on dispose uniquement d'une information "partielle" : on ne connaît pas la valeur de X mais on sait que $\mathbb{P}(X \notin A) \neq 0$.

Définition : Troncature à gauche
Soit Z une variable aléatoire indépendante de X, on dit qu'il y a troncature à gauche lorsque X n'est observable que si $X > Z$. On observe le couple (X, Z), avec $X > Z$.

Comme la troncature à gauche, on peut aussi définir la troncature à droite et la troncature par intervalle.

Dans le cadre de ce stage, les temps de décès des travailleurs de la cohorte EDF sont naturellement tronqués à gauche à la date d'entrée dans l'étude dès lors que l'échelle de temps considérée est l'âge (et non le temps de suivi). En efet, l'âge de décès ne peut être inférieur à l'âge d'entrée dans l'étude.

A.2 Théorème d'Hendry pour la simulation de données de survie avec covariables dépendantes du temps :

A.2.1 Preuve du théorème d'Hendry

On remarque que $\forall j \in \{1, \ldots, J\}$, la fonction de survie de Y est :

$$S_Y(t) = \frac{\exp\big(-\gamma_j(t - g^{-1}(s_{j-1}))\big)\prod_{h=1}^{j-1}\exp\big(-\gamma_h(g^{-1}(s_h) - g^{-1}(s_{h-1}))\big)}{K_E(g^{-1}(b)) - K_E(g^{-1}(a))} \times \mathbb{1}_{\{g^{-1}(s_{j-1}) < t \leq g^{-1}(s_j)\}}$$

Pour $X = g(Y)$, on a

$$\begin{aligned} S_X(t) &= \mathbb{P}(X > t) \\ &= \mathbb{P}(g(Y) > t) \\ &= \mathbb{P}(Y > g^{-1}(t)) \qquad \text{car g croissante et différentiable} \\ &= S_Y(g^{-1}(t)) \end{aligned}$$

Donc $\forall j \in \{1, \ldots, J\}$, la fonction de survie de X est définie par :

$$S_X(t) = \frac{\exp\big(-\gamma_j(g^{-1}(t) - g^{-1}(s_{j-1}))\big)\prod_{h=1}^{j-1}\exp\big(-\gamma_h(g^{-1}(s_h) - g^{-1}(s_{h-1}))\big)}{K_E(g^{-1}(b)) - K_E(g^{-1}(a))} \times \mathbb{1}_{\{g^{-1}(s_{j-1}) < g^{-1}(t) \leq g^{-1}(s_j)\}}$$

et sa densité de probabilité est définie par :

$$\begin{aligned} f_X(t) &= -\frac{\mathrm{d}S_X(t)}{\mathrm{d}t} \\ &= \frac{[\frac{\mathrm{d}}{\mathrm{d}t}g^{-1}(t)]\gamma_j\exp\big(-\gamma_j(g^{-1}(t) - g^{-1}(s_{j-1}))\big)\prod_{h=1}^{j-1}\exp\big(-\gamma_h(g^{-1}(s_h) - g^{-1}(s_{h-1}))\big)}{K_E(g^{-1}(b)) - K_E(g^{-1}(a))} \\ &\quad \times \mathbb{1}_{\{g^{-1}(s_{j-1}) < g^{-1}(t) \leq g^{-1}(s_j)\}} \end{aligned}$$

On en déduit que, pour tout $t \in]s_{j-1}, s_j]$, la fonction de risque instantané de X est définie par :

$$\begin{aligned} h_X(t) &= \frac{f_X(t)}{S_X(t)} \\ &= \frac{\mathrm{d}}{\mathrm{d}t} g^{-1}(t) \times \gamma_j \\ &= \frac{\mathrm{d}}{\mathrm{d}t} g^{-1}(t) \times \exp(\beta Z_j) \end{aligned}$$

En posant $h_0(t) = \frac{d}{dt}[g^{-1}(t)]$, on en déduit que X suit un modèle de Cox de risque instantané de base $h_0(t)$ et tronqué sur le support de temps [a,b].

A.2.2 Méthode d'acceptation-rejet

On voudrait simuler une variable aléatoire réelle X de densité de probabilité f. On suppose qu'il existe une autre densité de probabilité g telle que le ratio $\frac{f}{g}$ soit borné, disons par c (i.e. $f \leq cg$) et qu'on sache simuler Y de densité g Alors la méthode de rejet prend la forme suivante :

- Tirer Y de densité g ;
- Tirer U selon la loi uniforme $\mathcal{U}(0,1)$ indépendamment de Y ;
- Tant que $U > \frac{f(Y)}{cg(Y)}$ reprendre au premier point, sinon accepter Y comme un tirage alétoire de densité de probabilité f.

A.3 Vraisemblance des modèles de Cox et en EHR avec covariables temps-dépendantes et pour un taux de base constant par morceaux

Pour le calcul de la puissance statistique relative à la cohorte EDF il faut simuler des données. Afin de simuler des données selon les modèles de Cox et EHR, il est impératif de paramétrer complètement ces modèles, c'est-à-dire de donner une expression au risque de base. Une solution est de choisir h_0 constant par morceaux :
$\forall t \in I_l =]c_{l-1}, c_l]$ et pour $l \in \{1, \ldots, L\}$

$$h_0(t) = \lambda_l$$

où $0 = c_0 < c_1 < \ldots < c_L$ est une partition donnée du temps.

Considérons une partition du temps (commune à tous les individus)
$0 = s_0 < s_1 < \ldots < s_J$ telle que :
$\forall j \in \{1, \ldots, J\}$ et $\forall t \in]s_{j-1}, s_j]$, la dose cumulée $D^{cum}(t-10)$, noté Z_j, est constante sur l'intervalle $]s_{j-1}, s_j]$.

Posons $s_t = \max\{s_j \leq t; j \in \{0, \ldots, J\}\}$, j_t l'indice de l'intervalle correspondant à s_t et $\delta_l(t) = \mathbb{1}_{t \in I_l}$. Avec ces notations, la probabilité $S_i(t_i)$ que le travailleur i survive jusqu'au temps t_i sachant son exposition cumulée $D_i^{cum}(t_i - 10)$ est :

$$\begin{aligned}
S_i(t_i) &= \exp\left(-\int_0^{t_i} h(u)du \right) \\
&= \exp\left(-\int_0^{t_i} h_0(u)\rho(D_i^{cum}(u-10), \beta)du \right) \\
&= \exp\Bigg(-\sum_{\substack{l \neq 0 \\ c_l \leq t_i}} \lambda_l \sum_{\substack{j \neq 0 \\ c_{l-1} < s_j \leq c_l}} (s_j - s_{j-1})\rho(Z_{ij}, \beta) - \sum_{l=1}^{L} \lambda_l \delta_l(t_i) \Big[(t_i - s_t)\rho(Z_{i,j_t}, \beta) \\
&\qquad + \sum_{\substack{j \neq 0 \\ c_{l-1} < s_j \leq t_i}} (s_j - s_{j-1})\rho(Z_{ij}, \beta) \Big] \Bigg)
\end{aligned}$$

Dans le cadre de ce stage, les temps de décès sont tronqués à gauche à l'âge d'entrée dans l'étude.
Notons r_i l'âge du travailleur i à l'entrée dans l'étude et $\tilde{S}_i(t)$ (respectivement $\tilde{f}_i(t)$) la survie (respectivement la densité) tronquée de l'individu i jusqu'au temps t (respectivement au temps t), alors la densité tronquée de l'individu i en t s'écrit :

$$\tilde{f}_i(t) = \frac{f_i(t)}{S_i(r_i)}$$

En effet, dans ce cas

$$\begin{aligned}
\int_0^{+\infty} f_i(u)du = 1 &\Leftrightarrow \int_{r_i}^{+\infty} f_i(u)du + \int_0^{r_i} f_i(u)du = 1 \\
&\Leftrightarrow \int_{r_i}^{+\infty} f_i(u)du = S_i(r_i) \\
&\Leftrightarrow \int_{r_i}^{+\infty} \frac{f_i(u)}{S_i(r_i)}du = 1
\end{aligned}$$

De même, la survie tronquée du travailleur i s'écrit :

$$\tilde{S}_i(t_i) = \frac{S_i(t_i)}{S_i(r_i)}$$

En posant $\lambda = (\lambda_1, \ldots, \lambda_J)$, la vraisemblance de l'individu i est :

$$\begin{aligned} L(t_i|\beta,\lambda) &= \mathbb{P}(T_i \in [t_i, t_i+dt], \delta_i = 1|\beta,\lambda)^{\delta_i} \times \mathbb{P}(T_i \in [t_i, t_i+dt], \delta_i = 0|\beta,\lambda)^{1-\delta_i} \\ &= \mathbb{P}(X_i \in [t_i, t_i+dt], C_i \geq X_i|\beta,\lambda)^{\delta_i} \times \mathbb{P}(C_i \in [t_i, t_i+dt], C_i < X_i|\beta,\lambda)^{1-\delta_i} \\ &= \left(\tilde{f}(t_i|\beta,\lambda)G(t_i^-)\right)^{\delta_i} \times \left(g(t_i)\tilde{S}(t_i|\beta,\lambda)\right)^{1-\delta_i} \end{aligned}$$

Où $\tilde{f}$ (respectivent g) est la densité de X (respectivement C) et $\tilde{S}$ (respectivement G) est la survie de X (respectivement C).
Le but étant de maximiser cette vraisemblance et vu que la loi de C ne dépend pas des paramètres du modèle, il suffit de maximiser :

$$L(t_i|\beta,\lambda) = \tilde{f}_i(t_i)^{\delta_i}\tilde{S}_i(t_i)^{1-\delta_i}$$

En posant $\delta = (\delta_1, \ldots, \delta_n)$, on déduit que la vraisemblance du modèle est :

$$\begin{aligned} L(t_1, t_2, \ldots, t_n|\delta,\beta,\lambda) &= \prod_{i=1}^{n} \tilde{f}_i(t_i)^{\delta_i}\tilde{S}_i(t_i)^{1-\delta_i} \\ &= \prod_{i=1}^{n} \left(\tilde{S}_i(t_i) \times h_i(t_i)\right)^{\delta_i}\tilde{S}_i(t_i)^{1-\delta_i} \\ &= \prod_{i=1}^{n} h_i(t_i)^{\delta_i} \times \tilde{S}_i(t_i) \\ &= \prod_{i=1}^{n} \left[h_0(t_i)\rho(Z_{i,j_{t_i}},\beta)\right]^{\delta_i} \times \tilde{S}_i(t_i) \end{aligned}$$

Enfin, la log-vraisemblance est :

$$\begin{aligned} l(t_1, t_2, \ldots, t_n|\delta,\beta,\lambda) &= \sum_{i=1}^{n} \Big[\delta_i \times \big[\ln h_0(t_i) + \ln \rho(Z_{i,j_{t_i}},\beta)\big] + \ln \tilde{S}_i(t_i)\Big] \\ &= \sum_{i=1}^{n} \Big[\delta_i \times \big[\ln h_0(t_i) + \ln \rho(Z_{i,j_{t_i}},\beta)\big] + \ln S_i(t_i) - \ln S_i(r_i)\Big] \end{aligned}$$

A.4 Gradient de la vraisemblance

Soit l_i la log-vraisemblance du travailleur i.

— Pour le modèle de Cox :

la dérivée de $\ln S_i(t)$ par rapport à β est :

$$\frac{\partial \ln S_i(t)}{\partial \beta} = -\sum_{\substack{l\neq 0 \\ c_l \leq t}} \lambda_l \sum_{\substack{j\neq 0 \\ c_{l-1}<s_j\leq c_l}} (s_j - s_{j-1}) Z_{ij} \exp(\beta Z_{ij}) - \sum_{l=1}^{L} \lambda_l \delta_l(t) \Big[(t-s_t) Z_{i,j_t} \times \exp(\beta Z_{i,j_t}) + \sum_{\substack{j\neq 0 \\ c_{l-1}<s_j\leq t}} (s_j - s_{j-1}) Z_{ij} \exp(\beta Z_{ij}) \Big]$$

la dérivée de $\ln S_i(t)$ par rapport à λ_l est :

$$\frac{\partial \ln S_i(t)}{\partial \lambda_l} = -\mathbb{1}_{\{t>c_l\}} \sum_{\substack{j\neq 0 \\ c_{l-1}<s_j\leq c_l}} (s_j - s_{j-1}) \exp(\beta Z_{ij}) - \delta_l(t) \Big[(t-s_t) \exp(\beta Z_{i,j_t}) + \sum_{\substack{j\neq 0 \\ c_{l-1}<s_j\leq t}} (s_j - s_{j-1}) \exp(\beta Z_{ij}) \Big]$$

donc

$$\nabla \ln l_i(t) = \left(\frac{\partial \ln l_i(t)}{\partial \beta}, \frac{\partial \ln l_i(t)}{\partial \lambda_1}, \ldots, \frac{\partial \ln l_i(t)}{\partial \lambda_L} \right)$$

où

$$\frac{\partial \ln l_i(t)}{\partial \beta} = \delta_i Z_{i,j_t} + \frac{\partial \ln S_i(t)}{\partial \beta} - \frac{\partial \ln S_i(r_i)}{\partial \beta}$$

et

$$\frac{\partial \ln l_i(t)}{\partial \lambda_l} = \delta_i \frac{\delta_l(t)}{\lambda_l} + \frac{\partial \ln S_i(t)}{\partial \lambda_l} - \frac{\partial \ln S_i(r_i)}{\partial \lambda_l}$$

— Pour le modèle en EHR :

la dérivée de $\ln S_i(t)$ par rapport à β est :

$$\frac{\partial \ln S_i(t)}{\partial \beta} = -\sum_{\substack{l\neq 0 \\ c_l \leq t}} \lambda_l \sum_{\substack{j\neq 0 \\ c_{l-1}<s_j\leq c_l}} (s_j - s_{j-1}) Z_{ij} - \sum_{l=1}^{L} \lambda_l \delta_l(t) \Big[(t - s_t) Z_{i,j_t} + \sum_{\substack{j\neq 0 \\ c_{l-1}<s_j\leq t}} (s_j - s_{j-1}) Z_{ij} \Big]$$

la dérivée de $\ln S_i(t)$ par rapport à λ_l est :

$$\frac{\partial \ln S_i(t)}{\partial \lambda_l} = -\mathbb{1}_{\{t>c_l\}} \sum_{\substack{j\neq 0 \\ c_{l-1}<s_j\leq c_l}} (s_j - s_{j-1})(1 + \beta Z_{ij}) - \delta_l(t) \Big[(t - s_t)(1 + \beta Z_{i,j_t}) + \sum_{\substack{j\neq 0 \\ c_{l-1}<s_j\leq t}} (s_j - s_{j-1})(1 + \beta Z_{ij}) \Big]$$

donc

$$\nabla \ln l_i(t) = \left(\frac{\partial \ln l_i(t)}{\partial \beta}, \frac{\partial \ln l_i(t)}{\partial \lambda_1}, \ldots, \frac{\partial \ln l_i(t)}{\partial \lambda_L} \right)$$

où

$$\frac{\partial \ln l_i(t)}{\partial \beta} = \delta_i \frac{Z_{i,j_t}}{1 + \beta Z_{i,j_t}} + \frac{\partial \ln S_i(t)}{\partial \beta} - \frac{\partial \ln S_i(r_i)}{\partial \beta}$$

et

$$\frac{\partial \ln l_i(t)}{\partial \lambda_l} = \delta_i \frac{\delta_l(t)}{\lambda_l} + \frac{\partial \ln S_i(t)}{\partial \lambda_l} - \frac{\partial \ln S_i(r_i)}{\partial \lambda_l}$$

Pour les deux modèles le gradient de la vraisemblance est :

$$\nabla \ln l = \nabla \sum_{i=1}^{n} \ln l_i = \sum_{i=1}^{n} \nabla \ln l_i$$

A.5 Hessienne de la vraisemblance

Dans cette sous-partie, on présentera la hessienne de la contribution à la log-vraisemblance du travailleur i. Celle-ci nous permettera de calculer la matrice d'information de Fisher nécessaire pour répondre à la question b), i.e., pour calibrer le protocole d'études.

Soit l_i la contribution à la log-vraisemblance du travailleur i.

— Pour le modèle de Cox :

la dérivée seconde de $\ln S_i(t)$ par rapport à β est :

$$\frac{\partial^2 \ln S_i(t)}{\partial \beta^2} = -\sum_{\substack{l\neq 0 \\ c_l \leq t}} \lambda_l \sum_{\substack{j\neq 0 \\ c_{l-1}<s_j\leq c_l}} (s_j - s_{j-1}) Z_{ij}^2 \exp(\beta Z_{ij}) - \sum_{l=1}^{L} \lambda_l \delta_l(t) \Big[(t-s_t) Z_{i,j_t}^2 \times \exp(\beta Z_{i,j_t}) + \sum_{\substack{j\neq 0 \\ c_{l-1}<s_j\leq t}} (s_j - s_{j-1}) Z_{ij}^2 \exp(\beta Z_{ij}) \Big]$$

donc la dérivée seconde de l_i par rapport à β est :

$$\frac{\partial^2 l_i(t)}{\partial \beta^2} = \frac{\partial^2 \ln S_i(t)}{\partial \beta^2} - \frac{\partial^2 \ln S_i(r_i)}{\partial \beta^2}$$

la dérivée seconde de $\ln S_i(t)$ par rapport à λ_l est :

$$\frac{\partial^2 \ln S_i(t)}{\partial \lambda_l^2} = 0$$

donc la dérivée seconde de l_i par rapport à λ_l

$$\frac{\partial^2 l_i(t)}{\partial \lambda_l^2} = -\frac{\delta_i}{\lambda_l^2} \delta_l(t)$$

la dérivée de $\ln S_i(t)$ par rapport à λ_l et β est :

$$\frac{\partial^2 \ln S_i(t)}{\partial \lambda_l \partial \beta} = \frac{\partial^2 \ln S_i(t)}{\partial \beta \partial \lambda_l} = -\mathbb{1}_{\{t>c_l\}} \sum_{\substack{j\neq 0 \\ c_{l-1}<s_j\leq c_l}} (s_j - s_{j-1}) Z_{ij} \exp(\beta Z_{ij}) - \delta_l(t) \Big[(t-s_t) Z_{i,j_t} \exp(\beta Z_{i,j_t}) + \sum_{\substack{j\neq 0 \\ c_{l-1}<s_j\leq t}} (s_j - s_{j-1}) Z_{ij} \exp(\beta Z_{ij}) \Big]$$

donc la dérivée de l_i par rapport à λ_l et β est :

$$\frac{\partial^2 l_i(t)}{\partial \lambda_l \partial \beta} = \frac{\partial^2 \ln S_i(t)}{\partial \lambda_l \partial \beta} - \frac{\partial^2 \ln S_i(r_i)}{\partial \lambda_l \partial \beta}$$

— Pour le modèle en EHR :

la dérivée seconde de $l_i(t)$ par rapport à β est :

$$\frac{\partial^2 l_i(t)}{\partial \beta^2} = -\delta_i \frac{Z_{i,j_t}^2}{(1+\beta Z_{i,j_t})^2}$$

la dérivée seconde de $l_i(t)$ par rapport à λ_l est :

$$\frac{\partial^2 l_i(t)}{\partial \lambda_l^2} = -\frac{\delta_i}{\lambda_l^2}\delta_l(t)$$

la dérivée de $\ln S_i(t)$ par rapport à λ_l et β est :

$$\frac{\partial^2 \ln S_i(t)}{\partial \lambda_l \partial \beta} = \frac{\partial^2 \ln S_i(t)}{\partial \beta \partial \lambda_l} = -\mathbb{1}_{\{t>c_l\}} \sum_{\substack{j\neq 0 \\ c_{l-1}<s_j\leq c_l}} (s_j - s_{j-1})Z_{ij} - \delta_l(t)\Big[(t-s_t)Z_{i,j_t} + \sum_{\substack{j\neq 0 \\ c_{l-1}<s_j\leq t}} (s_j - s_{j-1})Z_{ij}\Big]$$

donc

$$\frac{\partial^2 l_i(t)}{\partial \lambda_l \partial \beta} = \frac{\partial^2 l_i(t)}{\partial \beta \partial \lambda_l} = \frac{\partial^2 \ln S_i(t)}{\partial \beta \partial \lambda_l} - \frac{\partial^2 \ln S_i(r_i)}{\partial \beta \partial \lambda_l}$$

Donc la hessienne de la vraisemblance est :

$$H(l(t)) = \sum_{i=1}^{n} H(l_i(t)) = \sum_{i=1}^{n} \begin{bmatrix} \frac{\partial^2 l_i(t)}{\partial \beta^2} & \frac{\partial^2 l_i(t)}{\partial \beta \partial \lambda_1} & \cdots & \frac{\partial^2 l_i(t)}{\partial \beta \partial \lambda_L} \\ \frac{\partial^2 l_i(t)}{\partial \lambda_1 \partial \beta} & \frac{\partial^2 l_i(t)}{\partial \lambda_1^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 l_i(t)}{\partial \lambda_L \partial \beta} & 0 & \cdots & \frac{\partial^2 l_i(t)}{\partial \lambda_L^2} \end{bmatrix}$$

A.6 Calcul de la matrice d'information de Fisher

Ici, on va calculer la matrice de Fisher uniquement pour le modèle de Cox.
Afin de calculer cette matrice, il faut calculer la loi du temps de survie $T = \min(X, C)$ où X et C sont deux variables aléatoires indépendantes de densité $f(t)$ (respectivement $g(t)$) et de survie $S(t)$ (respectivement $G(t)$), on a :

$$\begin{aligned} S_T(t) = \mathbb{P}(T > t) &= \mathbb{P}(\min(X, C) > t) \\ &= \mathbb{P}(X > t \text{ et } C > t) \\ &= \mathbb{P}(X > t)\mathbb{P}(C > t) \\ &= S(t)G(t) \end{aligned}$$

donc

$$f_T(t) = -\frac{dS_T(t)}{dt} = f(t)G(t) + g(t)S(t)$$

Remarque :

La fonction f_T est bien une densité, car elle est positive, intégrable sur $\mathbb{R}^+$ et :

$$\begin{aligned}
\int_{\mathbb{R}^+} f_T(t)dt &= \int_{\mathbb{R}^+} (f(t)G(t) + g(t)S(t))dt \\
&= \int_{\mathbb{R}^+} f(t)G(t)dt + \int_{\mathbb{R}^+} g(t) \int_t^{+\infty} f(u)dudt \\
&= \int_{\mathbb{R}^+} f(t)G(t)dt + \int_{\mathbb{R}^+} g(t) \int_{\mathbb{R}^+} f(u)\mathbb{1}_{\{t\leq u\}}dudt \\
&= \int_{\mathbb{R}^+} f(t)G(t)dt + \int_{\mathbb{R}^+} f(u) \int_{\mathbb{R}^+} g(t)\mathbb{1}_{\{t\leq u\}}dtdu \\
&= \int_{\mathbb{R}^+} f(t)G(t)dt + \int_{\mathbb{R}^+} f(u) \int_0^u g(t)dtdu \\
&= \int_{\mathbb{R}^+} f(t)G(t)dt + \int_{\mathbb{R}^+} f(u)(1 - G(u))du \\
&= \int_{\mathbb{R}^+} f(t)G(t)dt + 1 - \int_{\mathbb{R}^+} f(u)G(u)du \\
&= 1
\end{aligned}$$

On suppose que C_i suit une loi uniforme entre r_i et s_J.

Posons $I_i = \mathbb{E}(H(l_i(t))$ la matrice de Fisher de l'individu i.

Les calculs des formules ci-dessus sont détaillés en annexe.

Pour $j \in \{2, \ldots, L+1\}$, on a

$$[I_i]_{jj} = \mathbb{E}\left[\frac{\partial^2 l_i(T)}{\partial \lambda_{j-1}^2}\right] = -\frac{\delta_i}{\lambda_{j-1}^2}\left(\mathbb{1}_{\{r_i\leq c_{j-2}\}}(S_T(c_{j-2}) - S_T(c_{j-1})) + \mathbb{1}_{\{r_i\in]c_{j-2},c_{j-1}]\}}(S_T(r_i) - S_T(c_{j-1}))\right)$$

où $S_T(t) = \tilde{S}(t)G(t) = \dfrac{S(t)G(t)}{S(r_i)}$

et

$$[I_i]_{11} = \mathbb{E}\left[\frac{\partial^2 l_i(T)}{\partial\beta^2}\right] = -\frac{1}{S(r_i)(s_J - r_i)}\sum_{j=r_i}^{s_J}\sum_{k=1}^{L}\delta_{jk}e^{-\lambda_k S_i(c_{k-1},s_{j-1})-\sum_{m=1}^{k-1}\lambda_m S_i(c_{m-1},c_m)}$$
$$\times\left[\sum_{m=1}^{k-1}\lambda_m\frac{\partial^2 S_i(c_{m-1},c_m)}{\partial\beta^2}\Big(\big((1+\lambda_k s_J)A_{ijk} - \lambda_k B_{ijk} + \lambda_k(1+\lambda_k s_J)\right.$$
$$\times Z_{ij}^2 e^{\beta Z_{ij}}B_{ijk} - Z_{ij}^2 e^{\beta Z_{ij}}\lambda_k C_{ijk} - s_{j-1}Z_{ij}^2 e^{\beta Z_{ij}}\big((1+\lambda_k s_J)A_{ijk} - \lambda_k B_{ijk}\big)$$
$$\left.+\frac{\partial^2 S_i(c_{k-1},s_{j-1})}{\partial\beta^2}\big((1+\lambda_k s_J)A_{ijk} - \lambda_k B_{ijk}\big)\Big)\right]$$

où

$$\delta_{jk} = \mathbb{1}_{]s_{j-1},s_j]\subseteq]c_{k-1},c_k]}$$
$$S_i(x,y) = \sum_{\substack{h\neq 0\\ x<s_h\leq y}}(s_h - s_{h-1})e^{\beta Z_{ih}}$$
$$A_{ijk} = \int_{s_{j-1}}^{s_j} e^{-\lambda_k(t-s_{j-1})e^{\beta Z_{ij}}}dt = -\frac{1}{\lambda_k e^{\beta Z_{ij}}}(e^{\lambda_k(s_j-s_{j-1})e^{\beta Z_{ij}}} - 1)$$
$$B_{ijk} = \int_{s_{j-1}}^{s_j} te^{-\lambda_k(t-s_{j-1})e^{\beta Z_{ij}}}dt = -\frac{s_j e^{\lambda_k(s_j-s_{j-1})e^{\beta Z_{ij}}}}{\lambda_k e^{\beta Z_{ij}}} + \frac{s_{j-1}}{\lambda_k e^{\beta Z_{ij}}} + \frac{1}{\lambda_k e^{\beta Z_{ij}}}A_{ijk}$$
$$C_{ijk} = \int_{s_{j-1}}^{s_j} t^2 e^{-\lambda_k(t-s_{j-1})e^{\beta Z_{ij}}}dt = -\frac{s_j^2 e^{\lambda_k(s_j-s_{j-1})e^{\beta Z_{ij}}}}{\lambda_k e^{\beta Z_{ij}}} + \frac{s_{j-1}^2}{\lambda_k e^{\beta Z_{ij}}} + \frac{2}{\lambda_k e^{\beta Z_{ij}}}B_{ijk}$$

A.6.1 Écriture de la matrice de d'information de Fisher en fonction de Δn pour la calibration du protocole d'études

Afin de pouvoir optimiser ϕ qur $\mathcal{D}_Z$, il est nécessaire d'écrire la matrice d'information de Fisher en fonction de $\Delta n \in \mathcal{D}_Z$.

Écriture de la hessienne pour le modèle de Cox :

Soient Δn le nombre d'individus nouveaux et Δt le temps de suivi ajouté.

On sait que le temps de survie $T_i = min(X_i, C_i)$ de l'individu i est toujours inférieure à la censure déterministe à la date de point. Notons v_i l'âge de l'individu i à la date de point en 2003.

En utilisant les notations du chapitre 3, on peut écrire la hessienne de la vraisemblance du modèle de Cox sous la forme :

— par rapport à β :

$$\begin{aligned}\frac{\partial^2 l}{\partial \beta^2} &= -\sum_{i=1}^{n+\Delta n}\sum_{c_l \le v_i+\Delta t} \lambda_l \delta_l(t_i)\left[\sum_{j=1}^{v_i+\Delta t} \delta_j(t_i)[Z_{ij}^2(t_i - s_{j-1})e^{\beta Z_{ij}} + \frac{\partial^2}{\partial \beta^2} S_i(s_1, s_{j-1})]\right] \\ &= -\sum_{i=1}^{n}\sum_{l=1}^{v_i} \lambda_l \delta_l(t_i)\left[\sum_{j=1}^{v_i} E_{ij}(t_i)\right] - \sum_{i=1}^{n}\sum_{l=v_i+1}^{v_i+\Delta t} \lambda_l \delta_l(t_i)\left[\sum_{j=v_i+1}^{v_i+\Delta t} E_{ij}(t_i)\right] \\ &\quad - \sum_{i=n+1}^{n+\Delta n}\sum_{l=v_i+1}^{v_i+\Delta t} \lambda_l \delta_l(t_i)\left[\sum_{j=v_i+1}^{v_i+\Delta t} E_{ij}(t_i)\right] \\ &= C - \Psi_\beta(\Delta t) - \Lambda_\beta(\Delta n, \Delta t)\end{aligned}$$

où C est une contante par rapport à Δn et Δt et

$$E_{ij}(t_i) = \delta_j(t_i)[Z_{ij}^2(t_i - s_{j-1})e^{\beta Z_{ij}} + \frac{\partial^2}{\partial \beta^2} S_i(s_1, s_{j-1})]$$

Remarque :

Dans la formule ci-dessus, plusieurs termes s'annulent car on n'observe les temps de survie des nouveaux individus qu'à partir de la date de point en 2003, de plus

$$\sum_{l=v_i+1}^{v_i+\Delta t} \lambda_l \delta_l(t_i)\left[\sum_{j=1}^{v_i} E_{ij}(t_i)\right] = \sum_{l=1}^{v_i} \lambda_l \delta_l(t_i)\left[\sum_{j=v_i+1}^{v_i+\Delta t} E_{ij}(t_i)\right] = 0$$

— par rapport à λ_l :

$$\begin{aligned}\frac{\partial^2 l}{\partial \lambda_l^2} &= -\frac{1}{\lambda_l^2}\sum_{i=1}^{n+\Delta_n} \delta_i \delta_l(t_i) \mathbb{1}_{\{c_{l-1} < v_i + \Delta t\}} \\ &= -\frac{1}{\lambda_l^2}\sum_{i=1}^{n} \delta_i \delta_l(t_i) \mathbb{1}_{\{c_{l-1} < v_i + \Delta t\}} - \frac{1}{\lambda_l^2}\sum_{i=n+1}^{n+\Delta n} \delta_i \delta_l(t_i) \mathbb{1}_{\{c_{l-1} < v_i + \Delta t\}} \\ &= -\Psi_{\lambda_l}(\Delta t) - \Lambda_{\lambda_l}(\Delta n, \Delta t)\end{aligned}$$

— par rapport à λ_l et β :

$$\begin{aligned}\frac{\partial^2 l}{\partial \lambda_l \partial \beta} = -\sum_{i=1}^{n+\Delta n}\Bigg[&\mathbb{1}_{t_i > c_l}\mathbb{1}_{c_l \le v_i+\Delta t}\frac{\partial S_i(c_{l-1}, c_l)}{\partial \beta} + \delta_l(t_i)\mathbb{1}_{c_{l-1} \le v_i+\Delta t}\Bigg[\sum_{j=1}^{v_i+\Delta t} \delta_j(t_i)\Big[(t_i - s_{j-1}) \\ &\times Z_{ij}e^{\beta Z_{ij}} + \mathbb{1}_{c_l \le v_i+\Delta t}\frac{\partial S_i(c_{l-1}, c_l)}{\partial \beta} + \mathbb{1}_{c_l > v_i+\Delta t}\frac{\partial S_i(c_{l-1}, v_i + \Delta t)}{\partial \beta}\Big]\Bigg]\Bigg] \\ = -\Psi_{\lambda_l,\beta}(\Delta t) &- \Lambda_{\lambda_l,\beta}(\Delta n, \Delta t)\end{aligned}$$

où

$$\Psi_{\lambda_l,\beta}(\Delta t) = -\sum_{i=1}^{n}\Bigg[\mathbb{1}_{t_i>c_l}\mathbb{1}_{c_l\leq v_i+\Delta t}\frac{\partial S_i(c_{l-1},c_l)}{\partial\beta} + \delta_l(t_i)\mathbb{1}_{c_{l-1}\leq v_i+\Delta t}\Bigg[\sum_{j=1}^{v_i+\Delta t}\delta_j(t_i)\Big[(t_i-s_{j-1}) \times Z_{ij}e^{\beta Z_{ij}} + \mathbb{1}_{c_l\leq v_i+\Delta t}\frac{\partial S_i(c_{l-1},c_l)}{\partial\beta} + \mathbb{1}_{c_l>v_i+\Delta t}\frac{\partial S_i(c_{l-1},v_i+\Delta t)}{\partial\beta}\Big]\Bigg]\Bigg]$$

$$\Lambda_{\lambda_l,\beta}(\Delta n,\Delta t) = -\sum_{i=n+1}^{n+\Delta n}\Bigg[\mathbb{1}_{t_i>c_l}\mathbb{1}_{c_l\leq v_i+\Delta t}\frac{\partial S_i(c_{l-1},c_l)}{\partial\beta} + \delta_l(t_i)\mathbb{1}_{c_{l-1}\leq v_i+\Delta t}\Bigg[\sum_{j=1}^{v_i+\Delta t}\delta_j(t_i) \times \Big[(t_i-s_{j-1})Z_{ij}e^{\beta Z_{ij}} + \mathbb{1}_{c_l\leq v_i+\Delta t}\frac{\partial S_i(c_{l-1},c_l)}{\partial\beta} + \mathbb{1}_{c_l>v_i+\Delta t} \times \frac{\partial S_i(c_{l-1},v_i+\Delta t)}{\partial\beta}\Big]\Bigg]\Bigg]$$

Écriture de la matrice de Fisher :

En intégrant les relations précédentes par rapport à la loi de T, d'après les formules obtenues ci-dessus et en supposant que C_i suit une uniforme entre r_i et $s_J + \Delta t$, on obtient :

$$I(\hat{\lambda}_l,\Delta n) = -\frac{1}{\lambda_l^2}\sum_{i=1}^{n+\Delta n}\delta_i\Bigg(\mathbb{1}_{c_l\leq v_i+\Delta t}\big[\mathbb{1}_{r_i\leq c_{l-1}}(S_T(c_{l-1})-S_T(c_l)) + \mathbb{1}_{r_i\in]c_{l-1},c_l]}(S_T(r_i)-S_T(c_l))\big] + \mathbb{1}_{c_{l-1}\leq v_i+\Delta t}\Big[\mathbb{1}_{c_l>v_i+\Delta t}\big[\mathbb{1}_{r_i\leq c_{l-1}}(S_T(c_{l-1})-S_T(v_i+\Delta t)) + \mathbb{1}_{r_i\in]c_{l-1},v_i+\Delta t]} \times (S_T(r_i)-S_T(v_i+\Delta t))\big]\Big]\Bigg)$$

Comme pour la hessienne, la matrice de Fisher attendue par rapport à $\hat{\lambda}_l$ peut se décomposer en deux sommes ; une correspondante à la contribution des anciens invividus suivis plus longtemps, et une à la contribution des nouveaux individus suivi entre la date de point et la date de point plus un certain temps de suivi.

$$I(\hat{\beta},\Delta n) = -\sum_{i=1}^{n+\Delta n}\frac{1}{S(r_i)(s_J-r_i+\Delta t)}\sum_{j=r_i}^{v_i+\Delta t}\sum_{\substack{k\neq 0\\ 0<c_k\leq v_i+\Delta t}}\delta_{jk}e^{-\lambda_k S_i(c_{k-1},s_{j-1})-\sum_{m=1}^{k-1}\lambda_m S_i(c_{m-1},c_m)}$$
$$\times\Bigg[\sum_{m=1}^{k-1}\lambda_m\frac{\partial^2 S_i(c_{m-1},c_m)}{\partial\beta^2}\Bigg((1+\lambda_k(s_J+\Delta_t))A_{ijk}-\lambda_k B_{ijk}+\lambda_k(1+\lambda_k(s_J+\Delta_t))$$
$$\times Z_{ij}^2 e^{\beta Z_{ij}}B_{ijk}-Z_{ij}^2e^{\beta Z_{ij}}\lambda_k C_{ijk}-s_{j-1}Z_{ij}^2e^{\beta Z_{ij}}\big((1+\lambda_k(s_J+\Delta_t))A_{ijk}-\lambda_k B_{ijk}\big)$$
$$+\frac{\partial^2 S_i(c_{k-1},s_{j-1})}{\partial\beta^2}\big((1+\lambda_k(s_J+\Delta_t))A_{ijk}-\lambda_k B_{ijk}\big)\Bigg)\Bigg]$$

Cette somme peut se décomposer en trois sommes, une constante par rapport à Δn et Δt, une qui dépend des anciens individus suivis plus longtemps et une dernière qui dépend uniquement des nouveaux individus suivis entre la date de point et la date de point plus un certain temps de suivi.

$$I(\hat{\beta},\hat{\lambda}_l,\Delta n) = -\sum_{i=1}^{n+\Delta n}\frac{1}{S(r_i)(s_J-r_i+\Delta t)}\Bigg(\mathbb{1}_{c_l\leq v_i+\Delta t}\big(\mathbb{1}_{r_i\leq c_{l-1}}I(c_{l-1},c_l)+\mathbb{1}_{r_i\in]c_{l-1},c_l]}I(r_i,c_l)\big)$$
$$+\mathbb{1}_{c_l>v_i+\Delta t}\big(\mathbb{1}_{r_i\leq c_{l-1}}I(c_{l-1},v_i+\Delta t)+\mathbb{1}_{r_i\in]c_{l-1},c_l]}I(r_i,v_i+\Delta t)\big)$$
$$+\mathbb{1}_{r_i\leq c_l}J(c_l,vi+\Delta t)+\mathbb{1}_{r_i>c_l}J(r_i,v_i+\Delta t)\Bigg)$$

où

$$I(x,y) = \sum_{\substack{j\neq 0\\ x<s_j\leq y}}e^{-\lambda_l S_i(x,s_{j-1})-\sum_{c_k\leq x}\lambda_k S_i(c_{k-1},c_k)}\Bigg[Z_{ij}e^{\beta Z_{ij}}(1+(s_J+\Delta_t)\lambda_l e^{\beta Z_{ij}})(B_{ijl}-s_{j-1}A_{ijl})$$
$$-\lambda_l Z_{ij}e^{2\beta Z_{ij}}(C_{ijl}+s_{j-1}B_{ijl})+\frac{\partial S_i(x,s_{j-1})}{\partial\beta}\Big[(1+(s_J+\Delta_t)\lambda_l e^{\beta Z_{ij}}A_{ijl}-\lambda_l e^{\beta Z_{ij}}B_{ijl}\Big]\Bigg]$$

et

$$J(x,y) = \sum_{\substack{j\neq 0\\ x<s_j\leq y}}\sum_{k>l_x}\mathbb{1}_{]s_{j-1},s_j]\subset]c_{k-1},c_k]}e^{-\lambda_k S_i(c_{k-1},s_{j-1})-\sum_{c_m\leq c_{k-1}}\lambda_m S_i(c_{m-1},c_m)}$$
$$\times\big((1+\lambda_k(s_J+\Delta_t)e^{\beta Z_{ij}})A_{ijk}-\lambda_k e^{\beta Z_{ij}}B_{ijk}\big)$$

où $l_x = l$ si $x\in]c_{l-1},c_l]$.

A.7 Quelques résultats de puissance supplémentaires

A.7.1 Table et graphique de puissance pour le modèle en EHR à la date de point en 2003

TABLE A.1: Puissance statistique estimée au niveau $\alpha = 0.05$ pour la cohorte EDF initiale i.e., P_{2003} (date de point : 2003) pour la mise en évidence de différentes valeurs $exp(\beta)$ de ratio (pour 1 mSv) de risques instantanés de décès par cancer solide radio-induit à partir d'un modèle en EHR.

$exp(\beta)$	1.0001	1.0005	1.0009	1.001	1.003	1.005	1.007	1.009	1.015
P_{2003}	1.8%	0.6%	2%	2%	22 %	52%	79 %	97 %	100%

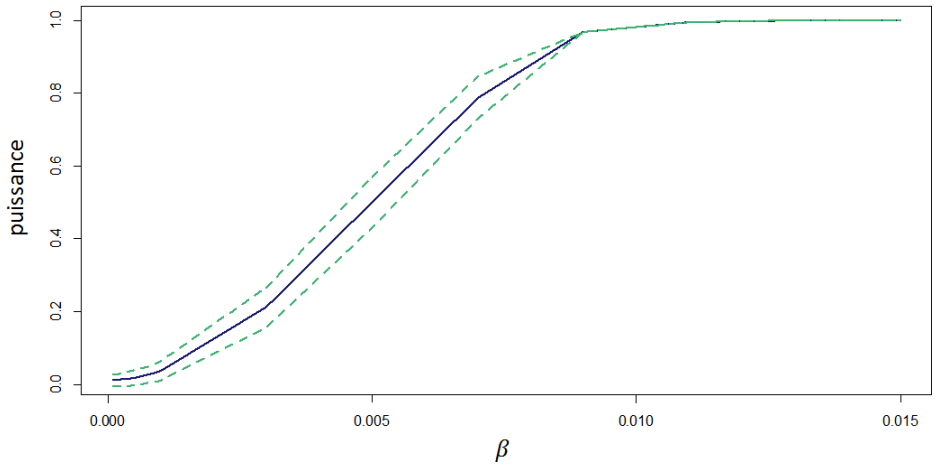

FIGURE A.2: Courbe de puissance statistique estimée en fonction du coefficient de risque β (ligne continue bleue) et erreur de Monte-Carlo associée (pointillés verts) au niveau $\alpha = 0.05$ pour la cohorte EDF initiale (date de point : 2003) pour la mise en évidence d'un risque de décès par cancer solide radio-induit à partir d'un modèle en EHR

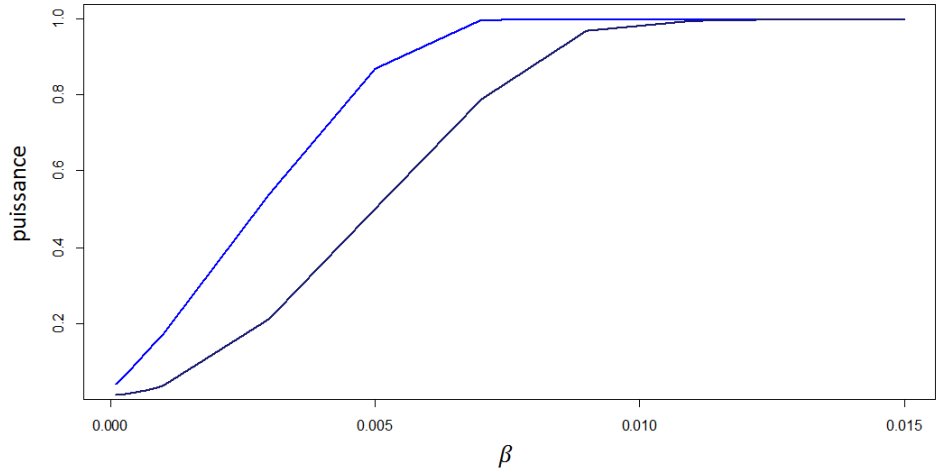

FIGURE A.3: Comparaison des courbes de puissance statistique approchée pour la cohorte EDF initiale (en noir) pour le modèle en EHR et pour le modèle de Cox (en bleue) en fonction du coefficient de risque β

Bibliographie

[1] S. Ancelet. Calibration bayésienne de plans d'expérience pour la quantification de sources d'incertitudes-application à l'estimation de la distribution statistique de la ténacité de l'acier de cuve. 2012.

[2] B. ARMOR. https://www.bellyarmor.fr/le-rayonnement-_r_27.html.

[3] A. Atkinson., A. Donev, and R. Tobias. Optimum experimental designs with sas. *Oxford Statistical Science Series*, 2007.

[4] X. Basagana and D. Spiegelman. The design of observational longitudinal studies.

[5] J. Bouyer and al. Epidémiologie : Principes et méthodes quantitatives. *Inserm*, 1996.

[6] E. Cardis and al. Risk of cancer after low doses of ionising radiation : retrospective cohort study in 15 countries. *BMJ*, 2005.

[7] B. Carlin and T. Louis. Bayesian methods for data analysis. *Chapman and Hall/CRC*, 2008.

[8] K. Chaloner and K. Larntz. Optimal bayesian design applied to logistic regression experiments. *Department of Applied Statistics, University of Minnesota, St. Paul, MN55108, U.S.A*, 1989.

[9] K. Chaloner and I. Verdinelli. Bayesian experimental design : a review. *Statistical Science*, 10(3) :273–304, 1995.

[10] D. Cox and D. Oakes. Analysis of survival data. *Chapman and Hall/CRC*, 1984.

[11] D. Hendry. Data generation for the cox proportional hazards model with time-dependant covariates : a method for medical researchers. *Statistics in Medicine*, 33(3) :436–454, 2013.

[12] S. Hoffmann, E. Rage, D. Laurier, P. Laroche, C. Guihenneuc, and S. Ancelet. Accounting for berkson and classical measurement error in radon exposure using a

bayesian structural approach in the analysis of lung cancer mortality in the french cohort of uranium miners. *Radiation Research Society*, 2017.

[13] J. Ibrahim, M. Chen, and D. Sinha. Bayesian survival analysis. *Springer*, 2001.

[14] ICRP. The 2007 recommendations of the international commission on radiological protection. *ICRP Publication 103. Ann. ICRP*, 37(2-4), 2007.

[15] Juhele. Penetrating power of different types of radiation - alpha, beta, gamma and neutrons. https://openclipart.org/detail/274074/penetrating-power-of-different-types-of-radiation-alpha-beta-gamma-and-\neutrons, 2017.

[16] J. Karvanen, J. Vanhatalo, K. Auranen, S. Kulathinal, and S. Mantyniemi. Optimal design of observational studies : overview and synthesis. 2017.

[17] O. Laurent, C. Metz, K. Joly, A. Rogel, and D. Laurier. Suivi et analyse épidémiologique des travailleurs d'Électricité de France surveillés pour exposition aux rayonnements ionisants, période 1961-2003. *IRSN*, 2011.

[18] O. Laurent, C. Metz-Flamant, A. Rogel, D. Hubert, A. Riedel, Y. Garcier, and D. Laurier. Relationship between occupational exposure to ionising radiation and mortality at the french electricity compagny, period 1961-2003. *Springer-Verlag. Int Arch Occup Environ Health*, 83 :935—-944, 2010.

[19] M. Little, R. Wakeford, J. Lubin, and G. M. Kendall. The statistical power of epidemiological studies analyzing the relationship between exposure to ionizing radiation and cancer, with special reference to childhood leukemia and natural background radiation. *Radiation Research*, 174(3) :387–402, 2010.

[20] K. Ozasa and al. Studies of the mortality of atomic bomb survivors, report 14, 1950–2003 : an overview of cancer and noncancer diseases. *Radiat. Res.*, 177 :229–243, 2012.

[21] M. Pearce and al. Radiation exposure from ct scans in childhood and subsequent risk of leukaemia and brain tumours : a retrospective cohort study. *The Lancet*, 380(9840) :499—-505, 2012.

[22] M. Portmann and A. Oulamara. Optimisation discrète. 2010.

[23] D. Richardson and al. Risk of cancer from occupational exposure to ionising radiation : retrospective cohort study of workers in france, the united kingdom, and the united states. *BMJ*, 351 :h5359, 2015.

[24] A. Rogel and al. Mortality of workers exposed to ionizing radiation at the french national electricity company. *American Journal of Industrial Medicine*, 47(1) :72–82, 2005.

[25] A. Rogel and al. Mortality in nuclear workers of the french electricity company : period 1968–2003. *Epidemiologie et Sante Publique*, 57(4) :e25–e33, 2009.

[26] P. Saint-Pierre. Processus de comptage et analyse de survie. 2015.

[27] D. Schoenfeld. Sample-size formula for the proportional-hazards regression model. *Biometrics*, 39(2), 1983.

[28] C. Scott and R. Nowak. A Neyman-Pearson approach to statistical learning. *IEEE transactions on information theory*, 51(11), 2005.

[29] T. Therneau and P. Grambsch. Modeling survival data : extending the cox model. *Springer-Verlag*, 2000.

[30] I. Thierry-Chef and al. The 15-country collaborative study of cancer risk among radiation workers in the nuclear industry : study of errors in dosimetry. *Radiation Research*, 167(4) :80–95, 2007.

[31] L. Thomas and E. Reyes. Tutorial : Survival estimation for cox regression models with time-varying coefficients using sas and r. *Journal of Statistical Software*, 2014.

[32] UNSCEAR. Unscear 2006 report to the general assembly with scientific annexes, effects of ionizing radiation. *UNSCEAR*, 2, 2008.

[33] A. Vazquez-Alcocer and al. Lades : A software for constructing and analyzing longitudinal designs in biomedical research. *PLoS ONE*, 9(7) :e100570, 2014.

[34] I. Verdinelli and J. Kadane. Bayesian design for maximising information and outcome. *Journal of the American Statistical Association*, 87(418) :510–515, 1992.

[35] I. Verdinelli, N. Polson, and N. Singpurwalla. Shannon information and bayesian design for prediction in accelereted life testing. *Chapman and Hall*, 1993.

[36] M. Zhou. Understanding the cox regression models with time-change covariates. *The American Statistician*, 55(2), 2001.